目　　录

第1章 MATLAB数据处理入门

在统计应用过程中，使用计算机进行数据处理是一种必然的选择．目前，已经有许多大型统计分析软件可供人们使用，最为著名的如 SAS, SPSS 和 SYSTAT 系统等．在这里我们选择 MATLAB 作为统计数据处理工具，不是因为它有更为强大的统计分析能力，而是因为 MATLAB 系统在科学计算领域的通用性，它更适合在专业众多的工科院校普及．

用 MATLAB 进行数据处理，可以在 MATLAB 的指令窗口(Command Window)中进行，也可以在 MATLAB Notebook 环境中进行(这是一种 MATLAB 与 Microsoft Word "无缝"链接的产物，是文字处理、数学计算和图形绘制一体化的工作环境)．本章简要介绍在 MATLAB 的工作环境下进行数据处理的基础知识．对于此前尚未接触过 MATLAB 的读者，本章内容可作为入门教程；对于已经了解 MATLAB 的读者，本章内容可作为课余阅读资料，教学中略过．

1.1 数值矩阵的建立与基本操作

1.1.1 数值矩阵的建立

MATLAB 语言主要的数据对象是数值矩阵．MATLAB 语言中，数值矩阵的输入方法有直接输入法、文件装载法、函数生成法．

1.1.1.1 直接输入法

直接输入法是由赋值语句实现的．赋值语句的基本结构是：

赋值变量＝赋值表达式

赋值变量通常是用户自定义变量，变量名是由英文字母引导的，由字母、数字和下划线组成；MATLAB 对字母的大小写是敏感的．

用赋值语句建立 m 行 n 列的二维数值矩阵(以下简称矩阵)的基本格式是：

矩阵＝[数据列表]

右侧的一对中括号"[ ]"是矩阵定义符，其中的数据列表排列成 m 行 n 列，每一行用分号"；"区分，行中元素用逗号"，"区分(逗号也可用一个空格代替)．

【例1.1】 建立一个 3×4 的矩阵 $\boldsymbol{A}$．

```
A=[1,2,3,4; 5,6,7,8; 9,10,11,12]
```

上述指令的运行结果是：

```
A =
     1     2     3     4
     5     6     7     8
     9    10    11    12
```

说明：输入必须在英文状态下进行．在定义符"[ ]"的结尾处(外侧)可以缀加一个分号";"，此时仅向 MATLAB 的工作内存输入了一个矩阵 **A**，但在当前的工作窗口中不输出(显示)这个矩阵．

在指令窗口中，输入一条指令后单击"Enter"键即可运行这一指令．在 Notebook 环境中，需要用鼠标选中这一指令，按住"Ctrl"键，然后单击"Enter"键即可运行这一指令．

【例 1.2】 当矩阵退化为一个数或一个向量时，可以由一个代数表达式加以定义，而不必使用矩阵定义符"[ ]"．

```
e=eps
r=2*(3+2*i)
P=(1:2:8)*pi
```

上述指令实际上是向 MATLAB 的工作内存输入了下列数值或向量：

```
e =
   2.2204e-016
r =
   6.0000 + 4.0000i
P =
   3.1416    9.4248    15.7080    21.9911
```

说明：① 表达式中的 eps，i，pi 都是 MATLAB 的保留常数．MATLAB 常用的保留常数与保留变量如下：

保留常数或变量	常数或变量的意义
eps	机器零阈值，2.2204×10^{-16}
i，j	虚数单位
pi	圆周率 π
Inf	$+\infty$的 MATLAB 表示
NaN	不定式 0/0 或∞/∞的 MATLAB 表示
ans	预定义缺省输出变量

这些不允许用户在自定义变量时使用．

② 最后一个赋值语句给出了向 MATLAB 输入一个等差数列的方法：在表达式

$a:d:b$ 中，a 是数列的首项，d 是数列的公差，b 是数列的上界(可能是最后一项或比最后一项大的数).

1.1.1.2 文件装载法

对于大规模的矩阵，通常预先编写数据文件存盘，然后使用“load”语句从数据文件中直接读入.

使用MATLAB系统自身的“内存变量编辑器”(Array Editor)编写数据文件十分方便可靠，操作方法是：

① 在指令窗口中向指定的新变量赋“空”矩阵，如 byk=[]；

② 在“内存变量浏览器”(Workspace)中双击该变量，启动“内存变量编辑器”；

③ 在“内存变量编辑器”弹出的空白表格中，每一个单元格对应矩阵的一个元素，填写具体数值；

④ 保存该变量为数据文件，如文件名为 byk，保存到 MATLAB 系统根目录下的 work 子目录中(或用户自己的工作目录下. 变量名与数据文件名可以不一致，一个数据文件也可以包含多个数据变量. 详细的内容请参阅其他 MATLAB 专门教程).

在需要调用这个数据文件时，只需运行指令 load byk 即可. 也可以利用任何一个文本编辑器(如 Microsoft Excel)编写这个数据文件，注意要保存为纯文本文件，如 byk.txt，在需要调用这个数据文件时，只需运行指令 load byk.txt.

1.1.1.3 函数生成法

在一些特殊的场合，需要用MATLAB定义的用来构造特殊矩阵的函数向系统输入数据. 使用MATLAB定义的函数称为**函数的调用**，这是MATLAB应用的一项重要技能. MATLAB函数调用语句的基本结构是：

[返回变量列表] = 函数名(输入变量列表)

其中，返回、输入变量列表中均可包含若干个变量，变量名之间用逗号分隔. 常用的构造特殊矩阵的函数有：

函数及调用格式	函数功能
Z=zeros(r,c)	生成元素全为0的r×c矩阵Z
O=ones(r,c)	生成元素全为1的r×c矩阵O
E=eye(r,c)	生成对角线为1、其他元素全为零的r×c矩阵E
D=diag(x)	生成以向量x的元素为对角元的对角矩阵D

在统计研究中，往往需要构造服从某一特定分布的随机数矩阵，这方面的函数较多，在后面的内容中会有这方面的应用，关于此类函数详细的内容请参阅本书附录B.

【例1.3】 特殊矩阵的函数生成.

```
Z = zeros(2,3) %生成2×3的全零矩阵
```

```
O=ones(3) %生成3阶全1方阵；方阵只需输入行数
E=eye(3,4) %生成3×4的对角线为1的矩阵
D=diag(1:5) %生成对角元为1、2、3、4、5的5阶对角矩阵
```

上述指令的运行结果是：

```
Z =
     0     0     0
     0     0     0
O =
     1     1     1
     1     1     1
     1     1     1
E =
     1     0     0     0
     0     1     0     0
     0     0     1     0
D =
     1     0     0     0     0
     0     2     0     0     0
     0     0     3     0     0
     0     0     0     4     0
     0     0     0     0     5
```

说明：① 表达式中的百分号“%”是 MATLAB 的注释符，用来对指令中的某些内容进行说明．“%”必须在英文状态下输入，其后的内容可以在中文状态下输入，指令运行时不执行这部分内容．

② diag 是一个双向操作函数，当输入参数 x 是一个向量时，输出(返回)以这个向量为对角元的对角矩阵；而当输入参数 x 是一个方阵时，则返回由这个方阵的对角元构成的列向量．

1.1.2 矩阵的基本操作

下面介绍关于矩阵元素的寻访与修改，以及矩阵的裁剪与拼接等矩阵操作方法．此类技能是使用 MATLAB 进行数据处理所必须的．

【例 1.4】 矩阵元素的寻访与修改．

```
A=[1,2,3,4; 2,3,4,5; 3,4,5,6; 4,5,6,7] %创建一个供操作的矩阵
A23=A(2,3) %寻访(取出)A的第2行、3列交叉位置的元素
```

```
A(2,2)=0 %将A的第2行、2列交叉位置元素赋值为零
```

上述指令的运行结果是：

```
A =
     1     2     3     4
     2     3     4     5
     3     4     5     6
     4     5     6     7
A23 =
     4
A =
     1     2     3     4
     2     0     4     5
     3     4     5     6
     4     5     6     7
```

【例1.5】 矩阵的裁剪(提取某些行、列，或删除某些行、列).

```
AR3=A(3,:) %取出A的第3行
AC2=A(:,2) %取出A的第2列
AR13=A(1:2:3,:) %取出A的第1、3两行
AR23C14=A(2:3,4:-3:1) %取出A的第2、3行与第4、1列交叉位置元素
A(:,4)=[] %删除A的第4列，矩阵的变量名不变
```

上述指令的运行结果是：

```
AR3 =
     3     4     5     6
AC2 =
     2
     0
     4
     5
AR13 =
     1     2     3     4
     3     4     5     6
AR23C14 =
     5     2
     6     3
```

```
A =
     1     2     3
     2     0     4
     3     4     5
     4     5     6
```

由例 1.5 可知，冒号“：”在 MATLAB 语言的矩阵操作中发挥着剪刀的作用.

【例 1.6】 矩阵的拼接(已知矩阵的扩展，或几个矩阵合并成一个新矩阵).

```
B=[A,ones(4,2)] % 在 A 的右边拼接 ones(4, 2)
C=[A(1:2,:);eye(3)] % 在 A 的 1、2 两行下边拼接 eye(3)
D=[A(1:2,2:3),zeros(2);ones(2,4)] % 在 A(1:2,2:3)右接 2 阶零矩阵，然后下接 2×4 全 1 矩阵
```

上述指令的运行结果是：

```
B =
    1     2     3     1     1
    2     0     4     1     1
    3     4     5     1     1
    4     5     6     1     1
C =
    1     2     3
    2     0     4
    1     0     0
    0     1     0
    0     0     1
D =
    2     3     0     0
    0     4     0     0
    1     1     1     1
    1     1     1     1
```

1.2 基本数学运算与常用函数

1.2.1 基本数学运算

MATLAB 数学运算的对象是矩阵，也就是说，要理解 MATLAB 中的数学运算，关

键是把握各种运算符作用于矩阵的规则.

1.2.1.1 矩阵的代数运算

MATLAB语言提供了如下矩阵代数运算的运算符:

'转置 +加法 -减法 *乘法 ^乘幂 \左除 /右除

上述运算符的运算规则遵循了线性代数教程中的相关定义. 矩阵代数运算要求维数相容, 否则将产生错误信息. 这里, 需要特别强调的是:

① 矩阵的转置“'”是指矩阵的共轭转置;

② 矩阵的左除“\”和右除“/”的含义是: 设 $\boldsymbol{A}$ 是可逆矩阵, 则 $\boldsymbol{AX}=\boldsymbol{B}$ 的解是 $\boldsymbol{A}$ 左除 $\boldsymbol{B}$, 即 $\boldsymbol{X}=\boldsymbol{A}\backslash\boldsymbol{B}$(若 $\boldsymbol{B}$ 为列向量, 则 $\boldsymbol{X}$ 为方程组的解); $\boldsymbol{XA}=\boldsymbol{B}$ 的解是 $\boldsymbol{A}$ 右除 $\boldsymbol{B}$, 即 $\boldsymbol{X}=\boldsymbol{B}/\boldsymbol{A}$(若 $\boldsymbol{B}$ 为列向量, 则 $\boldsymbol{X}$ 为方程组的解).

1.2.1.2 矩阵的标量批处理运算

MATLAB在需要的时候可以将矩阵视为普通的行列排列整齐的数据集合, 通常称为数组. 矩阵与数组在形式上是一样的, 但却是两个不同的概念. 当对一个矩阵(数组)施行标量批处理运算时, 这个矩阵就失去了线性代数教程中矩阵的意义而成为一个数组了. MATLAB语言提供了如下标量批处理运算的运算符:

.'转置 .*乘法 .^乘幂 .\左除 ./右除

上述运算符俗称“点运算”, 其运算规则是两个数组的对应元素之间的运算. 标量批处理运算要求数组的维数相同, 否则将产生错误信息. 特别地, 这里的转置“.'”是非共轭转置.

1.2.1.3 矩阵的关系运算

MATLAB语言提供了如表1.1所示的关系运算符.

表1.1 MATLAB语言提供的关系运算符

符号	<	>	<=	>=	==	~=
意义	小于	大于	不大于	不小于	等于	不等于
语法	A<B	A>B	A<=B	A>=B	A==B	A~=B

关系运算是在两个数值之间进行比较, 当给定的关系成立时返回数值1(表示关系真), 否则返回数值0(表示关系假). 当关系运算作用于一个标量与一个矩阵时, 是标量与矩阵的每一个元素进行比较, 返回一个与参与运算的矩阵同型的由0和1构成的矩阵; 当关系运算作用于两个同型矩阵时, 是两个矩阵的对应元素之间进行比较, 返回一个由0和1构成的同型矩阵.

1.2.1.4 矩阵的逻辑运算

MATLAB语言提供了如表1.2所示的逻辑运算符.

表 1.2　　　　MATLAB 语言提供的逻辑运算符

<table>
<tr><td colspan="2">符号</td><td>&</td><td>|</td><td>~</td></tr>
<tr><td colspan="2">意义</td><td>与</td><td>或</td><td>非</td></tr>
<tr><td colspan="2">语法</td><td rowspan="2">A & B</td><td rowspan="2">A | B</td><td rowspan="2">~A</td></tr>
<tr><td>A</td><td>B</td></tr>
<tr><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td></tr>
</table>

逻辑运算也是在两个数值之间进行的，运算过程中将任何非零元素视为 1(真). 当逻辑运算作用于一个标量与一个矩阵时，运算在标量与矩阵的每一个元素之间进行，返回一个与参与运算的矩阵同型的由 0 和 1 构成的矩阵；当逻辑运算作用于两个同型矩阵时，运算在两个矩阵的对应元素之间进行，返回一个由 0 和 1 构成的同型矩阵.

MATLAB 语言关于运算优先级的规定与数学中的规定是一致的.

【例 1.7】 两种转置运算的区别.

```
H=[(1:3)+(2:4)*i;1*i,-1*i,3] % 创建一个供操作的矩阵
H1=H'
H2=H.'
```

上述指令的运行结果是:

```
H =
    1.0000 + 2.0000i      2.0000 + 3.0000i      3.0000 + 4.0000i
         0 + 1.0000i           0 - 1.0000i      3.0000
H1 =
    1.0000 - 2.0000i           0 - 1.0000i
    2.0000 - 3.0000i           0 + 1.0000i
    3.0000 - 4.0000i      3.0000
H2 =
    1.0000 + 2.0000i           0 + 1.0000i
    2.0000 + 3.0000i           0 - 1.0000i
    3.0000 + 4.0000i      3.0000
```

【例 1.8】 两种乘、除和乘方运算的比较.

① 创建两个供操作的矩阵.

```
A=[1,2,3;0,1,2;0,0,1]
B=[0,0,1;0,2,1;3,2,1]
```

创建的两个矩阵是:

```
A =
      1     2     3
      0     1     2
      0     0     1
B =
      0     0     1
      0     2     1
      3     2     1
```

② 两种乘法运算的比较.

```
M1 = A * B
N1 = A. * B
```

这两个指令的运行结果是:

```
M1 =
      9    10     6
      6     6     3
      3     2     1
N1 =
      0     0     3
      0     2     2
      0     0     1
```

③ 两种除法运算的比较.

```
M2 = B/A
N2 = B./A
```

这两个指令的运行结果是:

```
M2 =
      0     0     1
      0     2    -3
      3    -4     0
Warning: Divide by zero.
N2 =
         0          0     0.3333
       NaN     2.0000     0.5000
       Inf        Inf     1.0000
```

④ 两种乘方运算的比较.

```
M3 = A^2
N3 = A.^2
```

这两个指令的运行结果是:

```
M3 =
     1     4    10
     0     1     4
     0     0     1
N3 =
     1     4     9
     0     1     4
     0     0     1
```

此外, 还应注意标量与矩阵运算的含义: 标量与矩阵的运算是标量与矩阵的每个元素之间的运算.

【例 1.9】 标量与矩阵的运算.

```
M4 = A * 10
N4 = A. * 10
```

这两个指令的运行结果是:

```
M4 =
    10    20    30
     0    10    20
     0     0    10
N4 =
    10    20    30
     0    10    20
     0     0    10
```

【思考题】 想一想, 下列运算的结果是什么?

① 5/A, 这个运算有意义吗?

② 5./B 和 B.\5, 这两种运算有意义吗? 运算的结果相同吗?

③ A./B 和 B.\A, 这两种运算的结果相同吗?

④ A/B 和 B\A, 这两种运算的结果相同吗?

【例 1.10】 标量与矩阵、矩阵与矩阵的关系运算.

```
x = 5
y = 5 * ones(3,3)
```

```
z=[1 2 3; 4 5 6; 7 8 10]
x~=z
y>=z
```

上述指令的运行结果是:

```
x =
     5
y =
     5     5     5
     5     5     5
     5     5     5
z =
     1     2     3
     4     5     6
     7     8    10
ans =
     1     1     1
     1     0     1
     1     1     1
ans =
     1     1     1
     1     1     0
     0     0     0
```

【例 1.11】 两个数组之间的逻辑运算.

```
a=1:9
b=9-a
c=~(a>4)
d=(a>=3)&(b<6)
```

上述指令的运行结果是:

```
a =
   1   2   3   4   5   6   7   8   9
b =
   8   7   6   5   4   3   2   1   0
c =
   1   1   1   1   0   0   0   0   0
```

```
d =
     0     0     0     1     1     1     1     1     1
```

1.2.2 统计数据处理常用的函数

MATLAB 提供了大量的函数，种类繁多. 按照函数的使用方法可以分为标量函数、向量函数和矩阵函数三种类型. 下面简要介绍这三种类型函数的一般概念，本书附录 A 列出了 MATLAB 核心程序包中的函数清单，需要时可通过 MATLAB 系统帮助进行学习.

1.2.2.1 标量函数

设 f 是 MATLAB 的标量函数，即对任意的 $X=(x_{ij})_{m\times n}$，有 $f(X)=(f(x_{ij}))_{m\times n}$. 标量函数的实质是矩阵元素的批处理运算，这些函数作用于矩阵时，是作用于矩阵的每一个元素(即函数的自变量实质上是矩阵的元素). 标量函数主要包含基本的数学函数，如三角函数、双曲函数、指数函数、对数函数、取整函数，等等. 统计数据处理中常用的标量函数有：

函数	功能描述
abs()	求绝对值
sqrt()	求平方根
exp(), pow2()	求以 e, 2 为底的指数
log10(), log(), log2()	求以 10, e, 2 为底的对数
sign()	符号函数
gamma()	Γ 函数
round()	四舍五入取整
ceil()	向 $+\infty$ 取整
floor()	向 $-\infty$ 取整
fix()	向 0 取整

对于各个函数的具体调用格式，除注意后面内容中的介绍之外，更细致的学习请运行指令

doc <函数名>

查询 MATLAB 系统帮助. 利用 MATLAB 系统的 Help，可以获得更多的帮助.

【例 1.12】 标量函数的功能.

```
x = [-1:0.25:0;0:0.25:1] % 创建一个自变量矩阵 x
xabs = abs(x) % 求矩阵 x 中元素的绝对值
xround = round(x) % 对矩阵 x 中元素进行四舍五入取整
```

```
xceil = ceil(x) % 对矩阵 x 中元素向 +∞取整
xexp = exp(x) % 计算 e^x 并把结果赋值给 xexp
```

上述指令的运行结果是：

```
x =
   -1.0000   -0.7500   -0.5000   -0.2500         0
         0    0.2500    0.5000    0.7500    1.0000
xabs =
    1.0000    0.7500    0.5000    0.2500         0
         0    0.2500    0.5000    0.7500    1.0000
xround =
    -1    -1    -1     0     0
     0     0     1     1     1
xceil =
    -1     0     0     0     0
     0     1     1     1     1
xexp =
    0.3679    0.4724    0.6065    0.7788    1.0000
    1.0000    1.2840    1.6487    2.1170    2.7183
```

1.2.2.2 向量函数

设 f 是 MATLAB 的向量函数，只有当其作用于向量 x 时才有意义(即函数的自变量是向量)．统计数据处理中常用的向量函数有：

函数	功能描述
sum()	求向量元素的和
cumsum()	求向量元素的累积和
prod()	求向量元素的积
cumprod()	求向量元素的累积积
max()	求向量元素的最大值
min()	求向量元素的最小值
sort()	对向量元素的排序操作
length()	查询向量的维数

向量函数也可以作用于矩阵，此时其作用对象是矩阵的列向量，运算的结果是一个行向量．

此外，各种统计量的计算函数也都是向量函数，这部分函数请注意后续内容中的介

绍，或参阅本书附录 B.

【例 1.13】 向量函数的功能.

```
A=[2,4,8,3;2,4,8,7;1,1,2,3;2,4,7,5;2,2,8,6]
xlength=length(A) %求 A 的列向量的长度
xsort=sort(A) %分别将 A 的各列元素从小到大排序
xsum=sum(A) %分别求 A 的各列元素的和
```

上述指令的运行结果是：

```
A =
     2     4     8     3
     2     4     8     7
     1     1     2     3
     2     4     7     5
     2     2     8     6
xlength =
     5
xsort =
     1     1     2     3
     2     2     7     3
     2     4     8     5
     2     4     8     6
     2     4     8     7
xsum =
     9    15    33    24
```

1.2.2.3 矩阵函数

设 f 是 MATLAB 的矩阵函数，即当 f 作用于矩阵 $\boldsymbol{A}$（即函数的自变量是矩阵）时，遵循线性代数中有关矩阵运算的规则．MATLAB 的矩阵函数按其作用可区分为：

① 矩阵生成与处理函数；

② 矩阵计算与线性方程组解算函数.

矩阵生成与处理函数在统计数据处理中比较常用，在 1.1.1 节对矩阵生成函数有部分介绍，这里再介绍几个矩阵处理函数：

函数	功能描述
reshape()	改变矩阵的型（将矩阵拉直为向量，或相反）
fliplr()	左右翻转矩阵

flipud()	上下翻转矩阵
rot90()	90°旋转矩阵
tril()	取矩阵的下三角部分
triu()	取矩阵的上三角部分

在矩阵计算与线性方程组解算函数中，统计数据处理常用的有：

函数	功能描述
size()	求矩阵的行与列的维数
det()	求矩阵的行列式
rank()	求矩阵的秩
norm()	求矩阵的范数
inv()	求矩阵的逆矩阵
pinv()	求矩阵的广义逆矩阵
eig()	求矩阵的特征值与特征向量
eigs()	求矩阵某些特定的(如最大)特征值及相应的特征向量

【例1.14】 矩阵函数的功能.

```
A=[2,4,8;2,8,7;1,2,3]
mn=size(A) % 求矩阵 A 的行与列的维数
ss=reshape(A,1,9) % 将矩阵 A 的各列首尾连接(拉直)成行向量 ss(1×9 矩阵)
AA=reshape(ss,3,3) % 用向量 ss 构造一个 3×3 矩阵 AA，注意与 A 比较
[V,D]=eigs(A) % 求矩阵 A 的特征值(D)及相应的单位正交特征向量(V)
```

上述指令的运行结果是：

```
A =
     2     4     8
     2     8     7
     1     2     3
mn =
     3     3
ss =
     2     2     1     4     8     2     8     7     3
AA =
     2     4     8
     2     8     7
     1     2     3
V =
```

```
    -0.5449      0.7955     -0.9561
    -0.8002     -0.5488     -0.0161
    -0.2507      0.2569      0.2924
D =
    11.5555           0           0
          0      1.8241           0
          0           0     -0.3795
```

1.3 数据图形化的常用指令与图形的简单修饰

在统计分析中将数据图形化，能够使数据所承载的信息生动、直观地展示出来. MATLAB 的语言具有十分强大的绘图功能，利用 MATLAB 语言的绘图指令可以简捷、自如地实现统计数据及分析结果的图形化. 这里我们仅对 MATLAB 语言数据图形化的基础知识作简单介绍.

1.3.1 数据图形化的常用指令

1.3.1.1 数据图形化的几个常用指令简介

MATLAB 语言的绘图指令十分丰富，这里仅介绍几个数据图形化最基本的指令.

(1) pie 指令

【调用格式】 **pie(y, explode)**

【功能】 绘制由数据向量 y(表示各个因素所占的百分比)所定义的圆饼图.

【参数说明】 参数 explode 是一个与 y 的维数相同的由数字 0 和 1 构成的向量，其作用是当其某个元素为 1 时，将对应的扇形从圆饼图形中分离出来.

参数 explode 可以缺省.

【扩展】 指令 pie3 绘制三维立体圆饼图，调用格式与 pie 相同.

(2) bar 指令

【调用格式】 **bar(x, y, 'option')**

【功能】 绘制以向量 x 的各个元素为横坐标，以向量 y 的各个对应元素为纵坐标所定义的条形图.

【参数说明】 向量 x 可以缺省，此时 x=1, 2, 3, ….

在 x 缺省时，参数 option 取值为 stacked 时绘制累加式条形图，以矩阵 y 的各个列向量的累加值为各矩形条的纵坐标；取值为 grouped 时绘制分组式条形图，以矩阵 y 的第 k 行数据为第 k 组中各矩形条的纵坐标.

参数 option 也可引用颜色参数，具体取值详见后面的"图形的简单修饰"．指定向量 x 时，option 只能引用颜色参数．

【扩展】 指令 barh 绘制水平放置的条形图，指令 bar3 绘制三维立体的垂直放置的条形图，指令 barh3 绘制三维立体的水平放置的条形图．调用格式均与 bar 类似．

(3) plot 指令

【调用格式】 `plot(x, y, 'option')`

【功能】 在二维直角坐标平面上绘制由向量 x 和 y 的对应元素为坐标的数据点或连接各个数据点的折线．

【参数说明】 向量 x 可以缺省，此时 x=1，2，3，…．

参数 option 的取值是表示线型、数据点标记、颜色的 1 个或几个符号，具体取值详见后面的"图形的简单修饰"．参数 option 可以缺省．

【扩展】 指令 plot3 绘制三维点线图，调用格式与 plot 类似．经常与 plot 指令配合使用的还有指令 line，其最简单的调用格式是 line(x, y)，功能是绘制出由向量 x 和 y 的对应坐标标记出的数据点(x_i，y_i)之间的折线，通常用 line 指令为 plot 图形中补充、添加辅助线．

指令 plot 和 line 还有其他更为细腻的调用格式，感兴趣的读者请查阅 MATLAB 系统帮助．

此外，MATLAB 语言还提供了很多更为专业的统计绘图指令，请注意后续内容中的介绍，或参阅本书附录 B．

1.3.1.2 多窗口绘图与点线图的单窗口多图方法

(1) 多窗口绘图方法

subplot 指令能够将当前的图形窗口分割成若干个子窗口，实现在每个子窗口分别绘制不同的图形的目的．指令的调用格式为

```
subplot(m, n, p)
```

其功能是将当前图形窗口分割成 m 行 n 列，并且现在正准备在第 p 个子窗口绘图．m 和 n 的最大取值是 9，即最多允许 9×9 的分割．子窗口的编号是从 1 至 m×n，先上后下，先左后右．撤销分割的方式是运行指令 clf 或 subplot(1, 1, 1)．

(2) 点线图的单窗口多图方法

点线图的单窗口多图方法是由 plot 指令的如下两种调用格式实现的：

```
plot(x, Y)
```

这是一种简捷的调用格式，矩阵 Y 的行(列)维数必须与向量 x 的长度相等；

```
plot(x1,y1,'option_1',x2,y2,'option_2', …,xn,yn,'option_n')
```

这是一种细腻的调用格式，向量 xk 与 yk 等长，k=1，2，…，n(不同的 k，向量长

度可以不相等)，参数 option_k 的取法同前.

这是一种批命令式的实现方式.

还有一种追加式的实现方式，要求 plot 指令与 hold on/off 指令配合使用，方法是：

```
plot(…)
hold on
plot(…)
hold off
```

这是因为，指令 plot 执行时首先对当前图形窗口清屏，然后绘制图形. 因此，在图形窗口只保留最新的 plot 图形. 在当前图形窗口中追加新 plot 图形，首先要由 hold on 指令通知系统保留当前图形窗口中的图形，hold off 指令的作用是解除 hold on 指令.

【例 1.15】 多窗口绘图方法(不同形式的条形图).

```
clear %清空工作内存
clf %重置当前图形窗口为初始状态
y1 = [15,35,10,20,20];
y2 = [15,35,10;20,20,15;10,15,30];
subplot(2,2,1)
bar(y1,'c') %简单条形图
subplot(2,2,2)
bar(y2,'grouped') %垂直放置的分组式条形图
subplot(2,2,3)
barh(y2,'stacked') %水平放置的累加式条形图
subplot(2,2,4)
bar3(y2,'grouped') %垂直放置的三维立体条形图
```

上述指令的运行结果见图 1.1.

1.3.2 图形的简单修饰

在绘图过程中，为使图形更加美观、易读，对图形进行一些简单的修饰是必要的.

1.3.2.1 点线图中的线型与数据点的标记，图形中的颜色

绘制点线图时，可以根据需要对线型、数据点标记及其颜色作出选择.

(1) 线型与参数取值

线型与参数取值见表 1.3.

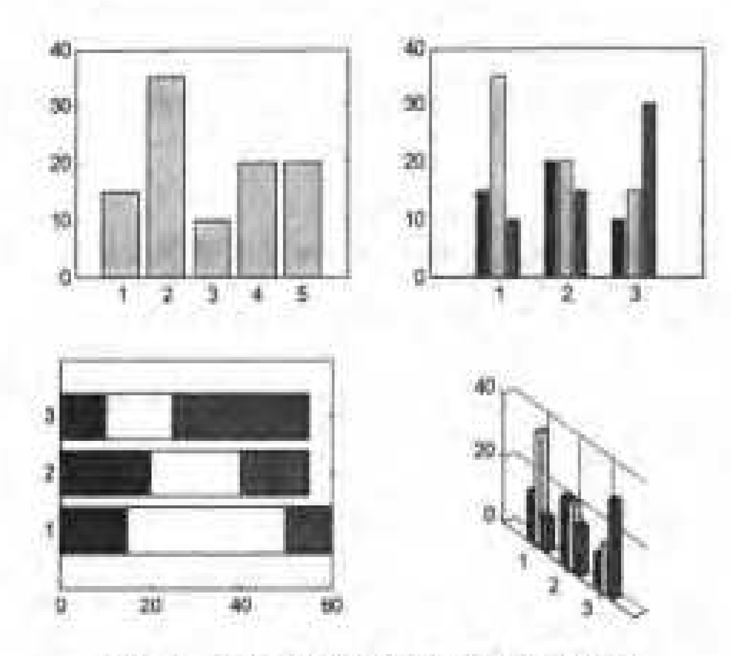


图 1.1 多窗口绘图示例(不同形式的条形图)

表 1.3 线型与参数取值表

线型	实线(默认)	点线	点划线	虚线
Option 值	-	:	-.	--

(2) 数据点形状与参数取值

数据点形状与参数取值见表 1.4.

表 1.4 数据点形状与参数取值表

数据点形状	实心点(默认)	加号形	雪花形	乘号形	空心圆形
Option 值	.	+	*	x	o
数据点形状	空心正方形	空心菱形	空心三角形	空心五角星	空心六角星
Option 值	s	d	^,v,>,<	p	h

(3) 颜色与参数取值

颜色与参数取值见表 1.5.

表 1.5 颜色与参数取值表

颜色	蓝(默认)	洋红	蓝绿	黄	红	绿	白	黑
Option 值	b	m	c	y	r	g	w	k

【例 1.16】 点线图的修饰与同一窗口多图画法.

```
clear, clf
x = 0:0.5:5; y = exp(x); % 创建指数函数数据
z = 0:pi/50: 2 * pi; F = [sin(z);cos(z);sin(2 * z)]; % 创建三组三角函数数据
```

```
subplot(3,1,1)
plot(x,y,'s-.m') % 用洋红色正方形绘制指数函数数据点，并用点划线连接数据点
subplot(3,1,2)
plot(z,F(1,:),z,F(2,:),'sg',z,F(3,:),'.c') % 绘制由矩阵 F 定义的三组三角函数的图形，分别指定数据点形状、线型和颜色
subplot(3,1,3)
plot(z,F) % 绘制由矩阵 F 定义的三组三角函数的图形，系统自动处理
```

上述指令的运行结果见图 1.2.

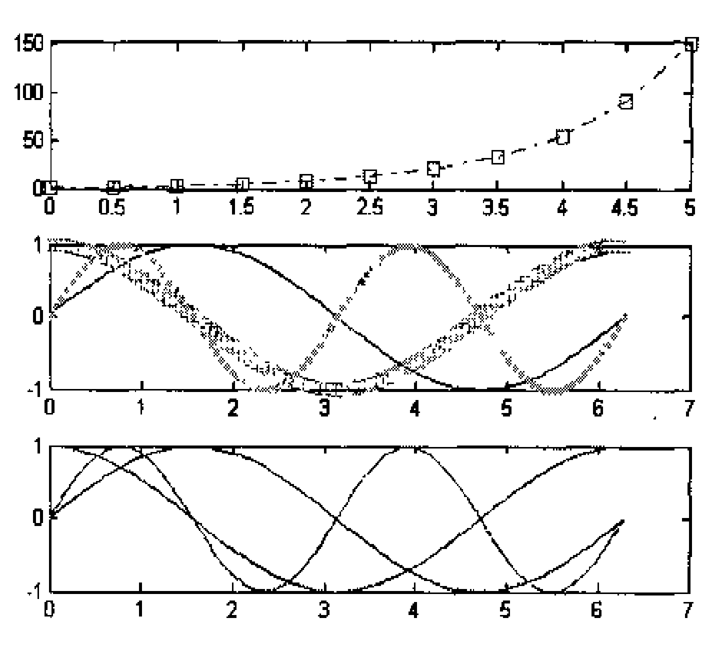


图 1.2 点线图的修饰与同一窗口多图画法示例

1.3.2.2 图形的标记

图形的标记主要包括下列内容：设置图形标题，设置坐标轴名称，添加图例，添加注释文字. 下面介绍相关指令.

(1) title 指令

【调用格式】 `title('string')`

【功能】 设置图形标题.

【说明】 在所画图形的最上端显示说明该图形标题的字符串 string.

(2) xlabel/ylabel 指令

【调用格式】

`xlabel('string')`

ylabel('string')

【功能】 设置坐标轴名称.

【说明】 xlabel('string')指令将字符串string水平放置于横轴下方，以说明横轴数据的意义；ylabel('string')指令将字符串string垂直放置于纵轴左侧，以说明纵轴数据的意义.

(3) legend指令

【调用格式】 legend('string1','string2','string3',…,option)

【功能】 添加图例.

【说明】 为图形按绘图的先后次序，用对应顺序的字符串string添加图例. 参数option可以省略，此时图例自动放置在图形视窗之内. 当option = -1时，表示强行将图例放置到图形视窗之外.

(4) text指令

【调用格式】 text(x,y,'string','cs')

【功能】 添加注释文字.

【说明】 在图形的指定坐标位置(x, y)处，添加由字符串string所给出的注释文字. cs是可选的引用参数，如果不给出该选项，则(x, y)坐标的度量单位与图形中数据单位一致；如果给出该选项，则(x, y)坐标表示规范化图形窗口的相对坐标，其变化范围是0~1的实数，图形窗口的左下角坐标为(0, 0)，右上角坐标为(1, 1).

在使用上述标记指令时，可以对字符串string所给出的文字字号的大小进行控制，其设置方法是：在字符串string所给出的注释文字的前面(单引号内)添加控制参数\fontsize{number}，number的取值为整数，缺省值为10.

【例1.17】 图形中标记的设置.

```
clear
clf
x = 0:pi/50:2 * pi;
Y = [sin(x);sin(2 * x);cos(x)];
plot(x,Y)
title('\fontsize{18}三角函数图像') %设置标题
xlabel('\fontsize{12}弧度值') %设置横轴说明
ylabel('\fontsize{16}函数值') %设置纵轴说明
legend('sin(x)', 'sin(2x)', 'cos(x)', -1) %设置图例
text(0.8,0.71,'\fontsize{12}←sin(x)和cos(x)在π/4的交点') %设置注释
```

上述指令的运行结果见图1.3.

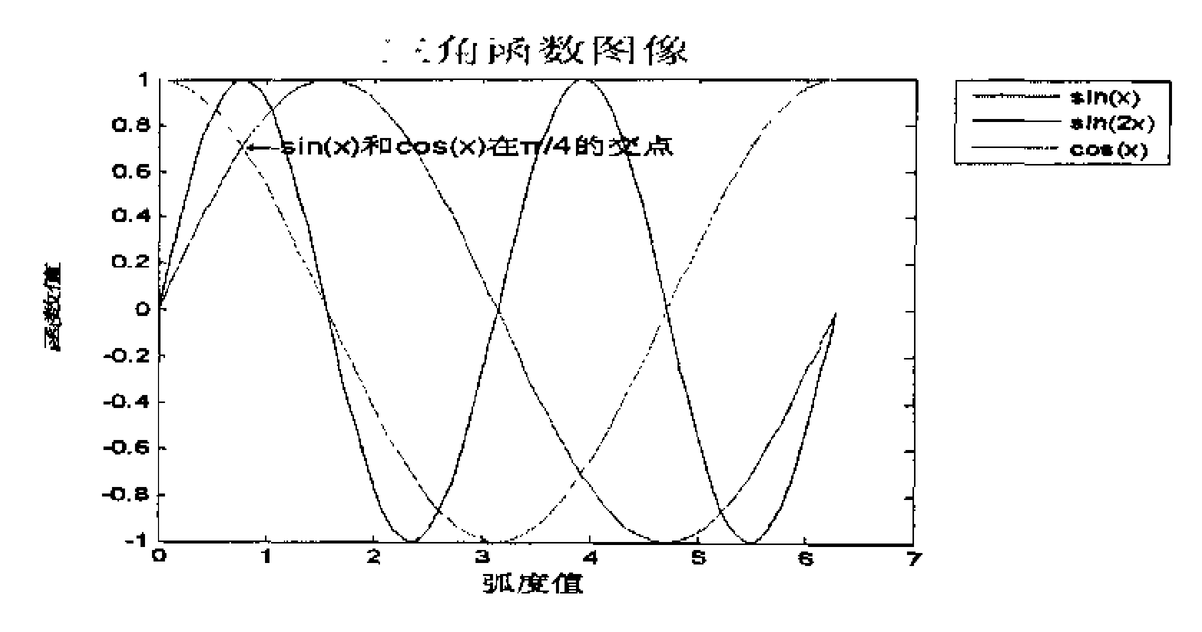


图 1.3　图形标记设置示例

1.4　运算流程的控制与指令集的函数化

1.4.1　运算流程的控制

(1) for-end 循环结构

【语法】

```
for i = a:k:b
        <commands>
end
```

【说明】 for 循环结构的执行与 C 语言、VB 语言相似. i 为循环变量, a 为循环初值, k 为步长, b 为循环终值; commands 为循环体.

(2) if-end 分支结构

【语法 Ⅰ】

```
if <expression>
        <commands>
end
```

【语法 Ⅱ】

```
if <expression>
        <commands_1>
```

```
else
        <commands_2>
end
```

【语法Ⅲ】

```
if <expression_1>
        <commands_1>
elseif <expression_2>
        <commands_2>
         ……
else
        <commands_k>
end
```

【说明】 分支结构的执行与C语言、VB语言相似. expression是关系或逻辑判断语句.

【例1.18】 运算流程的控制示例(求一个数的绝对值):

```
x=[-0.2, 0, 0.2];
xbas=[]; % 创建一个存放绝对值的动态数组(不指定维数的空向量)
for i=1:length(x)
    if x(i)>0
        xx=x(i); % 如果这个数是正数, 则不变号存入临时变量xx
    elseif x(i)<0
        xx=-x(i); % 否则, 如果这个数是负数, 则变号存入临时变量xx
    else
        xx=0; % 否则, 将0存入临时变量xx
    end
    xbas=[xbas,xx]; % 将当前临时变量xx中的数存入动态数组xbas
    end
    x,xbas % 显示数及其绝对值
```

上述指令的运行结果是:

```
x =
    -0.2000         0    0.2000
xbas =
     0.2000         0    0.2000
```

注意, MATLAB系统提供了求一个实数的绝对值(或复数的模)的计算指令abs.

MATLAB 系统还提供了其他几种控制程序流程的指令结构，包括 while-end 循环结构，switch-case-end 开关结构和 try-catch-end 探试结构．关于这几种指令结构的语法和使用技巧等，请读者参阅 MATLAB 系统帮助或其他 MATLAB 基础教程．

1.4.2 指令集的函数化

在前面的各个例子中，无论是计算还是绘图，我们都是根据问题的需要按照一定的顺序使用有关指令，这些指令构成了解决某一特定问题的指令集．可以将这类指令集保存为 MATLAB 语言的可执行文件，称为 M-脚本文件．这时，文件名就变成了一条可执行的(用户自定义)指令，以后若再次执行这一指令集，只需在指令窗口或 Notebook 环境中键入这个文件名并运行即可．M-脚本文件运行中所处理的数据及返回数据均在 MATLAB 的工作内存(Workspace)中．这种做法适宜小规模计算或编写大规模计算程序的主程序．

如果问题的规模较大，结构化程度高，相关算法在同类问题中可重复使用时，则应当对相应的 MATLAB 指令集由 function 指令进行函数化处理，规划和确定输入/输出参数，此类可执行文件称为 M-函数文件，通常用于子程序的编写．用户自定义的 M-函数同 MATLAB 系统函数一样，可以在需要时重复调用．与 M-脚本不同，M-函数运行中所处理的数据及返回数据，除预先定义的全局(输入/输出)变量在 MATLAB 的工作内存(Workspace)中，其余均在调用该函数时系统自动开辟的临时的局部变量空间中，该函数运行结束时系统自动删除这一临时的局部变量空间．

M-脚本文件与 M-函数文件是 MATLAB 语言程序设计的两种源程序文件格式，统称 M-文件，文件的扩展名均为 m.

M-文件的编写通常在 MATLAB 程序编辑器(Editor)中进行．单击“New M-File”图标(或菜单选项)，即可开启程序编辑器．若是在 Notebook 环境中已经编写出 M-文件，则可将文件中的全部指令复制/粘贴到程序编辑器，调试无错误即可保存．

下面简要介绍 M-文件的编写规范．

M-脚本文件和 M-函数文件的编写规范：除 M-函数文件必须要有由“function”引导的函数申明行外，其他要求一样．下面介绍 MATLAB 系统规范的 M-函数文件编写要求，遵循这一要求的用户程序文件可以纳入 MATLAB 系统进行管理．

函数申明行	function[返回变量列表]=funname(输入变量列表)
H1(关键词)行	%FUNNAME(大写体函数名)，关键词描述的函数功能
在线帮助文本区	%输入、输出变量的意义，调用格式说明；算法说明等
隔离行	无任何标记的空行
编写与修改记录	%编写者姓名，编写日期，修改日期等
隔离行	无任何标记的空行

函　　数　　体　MATLAB命令集(为增强程序的可读性，在函数体中可配置适当的空行和%引导的注释)

【例1.19】 M-函数文件的编写(改写例1.18中的指令集为M-函数文件，文件名为bykabs).

在例1.18的指令集中，明确向量x为输入参数，向量xabs为输出参数. 于是，由指令function定义M-函数文件bykabs的规范化过程是：

```
function xabs = bykabs(x)
%BYKABS函数的功能是求实数向量x的每一个元素的绝对值
%调用格式        xabs = bykabs(x)
%算法 当x>0时|x| = x, 当x = 0时|x| = 0, 当x<0时|x| = -x
% 输入参数  x是待求绝对值的实数向量
% 输出参数  xabs是向量x的绝对值向量


%包研科编写于2008年1月10日

xabs = [];
for i = 1:length(x)
    if x(i)>0
        xx = x(i);
    elseif x(i)<0
        xx = -x(i);
    else
        xx = 0;
    end
    xabs = [xabs,xx];
end
```

保存这个M-函数文件到用户自己的工作目录下或MATLAB系统根目录下的work子目录中，文件名为bykabs.m. 这样，M-函数文件bykabs就纳入了MATLAB系统的管理之中. 对于后来希望使用这个函数的用户，可以运行指令

```
doc bykabs
```

通过MATLAB系统的Help窗口，可以了解该M-函数文件的功能、调用方法、算法说明、参数意义. 使用时，只需按正确的调用格式调用这个函数即可. 如

```
clear
x = [-2,2,0,3.5,-2.3];
```

```
xabs = bykabs(x)
```

上述指令的运行结果是：

```
xabs =
        2.0000      2.0000       0      3.5000      2.3000
```

需要说明的是，此例简单到了只有一个输入参数和一个输出参数的情形．若有 m 个输入参数和 n 个输出参数，则函数申明行的形式为

```
function [y1,y2,…,yn] = FunName(x1,x2,…,xm)
```

参数的顺序按其重要性和是否可以缺省排列，重要的在前，可以缺省的靠后．

1.4.3 M-文件的保护

M-脚本文件和 M-函数文件均由 ASCⅡ码构成，可以由任何一种纯文本文件编辑器查看或修改文件的源代码．因此，为防止有人擅自修改这个文件，可用伪代码编译转化为二进制代码，不仅将文件保护起来，还可以提高程序的运行速度．

MATLAB 语言的伪编译生成的文件称为 P 代码文件，即文件的扩展名为 p.

生成 P 代码文件的指令是 pcode，其使用方法是：

```
pcode FunName - inplace
```

即在 M-函数文件 FunName.m 所在的目录上生成 FunName.p.

例如，前面我们将 bykabs.m 文件保存在当前工作路径"E:\数理统计\m 文件"下（即 MATLAB 系统窗口的"Current Directory"被设置为"E:\数理统计\m 文件"），则运行指令

```
pcode bykabs.m - inplace
```

就可在"E:\数理统计\m 文件"下生成一个 bykabs.p 文件，再调用函数 bykabs 时，系统运行的是 bykabs.p，而不是 bykabs.m.

当需要查看内存中 P 代码文件列表或清除内存中的 P 代码文件时，分别运行下列指令：

inmem 列出内存中所有的 P 代码文件

clear FunName 清除内存中 FunName.p 文件

clear functions 清除内存中所有的 P 代码文件

习题 1

1．写出下面 MATLAB 语句的运行结果（显示）：

```
A=[1,2,3;4,5,6;7,8,0];
A(1,:)*A(:,3)
sum(A.*A)
```

B=[A, ones(3,2)]

2．编写一个10×15矩阵的mat数据文件，保存到你的工作路径下，然后使用load指令读入这个矩阵.

3．写出下列语句的运行结果(显示)：

```
a=[1 2 3;4 5 6;7 8 9];
a1=a(2,:)
a2=a(:,2)
a3=a(:,3: -1:2)
a4=[a;a1]
```

4．把图形窗口分割成两个子窗口，分别画出下列曲线，试写出绘图指令集：

(1) 向量y1=[34,27,21,18]所定义的圆饼图，并将第2、第4块对应的扇形从圆饼图形中分离出来；

(2) 矩阵y2=[20,15,35;30,20,15;25,15,30]水平放置的累加式条形图.

5．在区间[0,2*pi]画sin(x)的图形，并添加图形标记“自变量x”“函数y”“sin(x)的图像”.

6．设x=[0:0.1:2*pi]，在同一窗口依次画出$y_1=\sin(x^2)$，$y_2=\sin(x)\cos(x)$的图像．要求：第一条曲线为红色点线，第二条曲线为蓝绿色虚线．试写出绘图指令集，并用你可以想到的方法在上述图形中加入各种解释和说明的文字.

7．编写M-脚本文件：对$n=1,2,\cdots,10$，求$x_n=\cos\left(\frac{n\pi}{10}+n\right)$的值.

8．编写M-函数文件：对任意$x_1,x_2\in\mathbf{R}$，求$f(x_1,x_2)=100(x_2-x_1^2)^2+(1-x_1)^2$的函数．调用这个函数，求$f(2,3)$的值.

9．编写M-函数文件：输入一个$m\times n$矩阵，返回矩阵元素的最大值及其所处的位置，然后随机生成一个3×4矩阵验证函数的功能.

第 2 章 统计分析的基本概念、工具与推理基础

变量与数据是统计分析的基本对象，统计量是统计分析的基础性工具，而抽样分布则构成统计推理的理论基础．本章简要介绍变量与数据、统计量、抽样分布的基本内容以及相关的 MATLAB 函数．

2.1 变量与数据的基本概念

2.1.1 变量及其概率分布

任何一个统计问题都要有明确的研究对象，称研究对象的全体为总体，每个具体的对象称为总体中的个体．

例如，在研究某大学一年级大学生的身体素质时，全体大学一年级新生就是总体．然而，人们用统计的方法研究一年级大学生的身体素质，关心的往往是每个学生(个体)的某项反映人的身体素质的指标，特别是可数量化的指标，如肺活量，关心这一数量化指标的指标值在群体中出现的规律．这个数量化指标才是统计研究的真正总体，通常称之为统计总体．

由此可以看出，总体的概念具有两重性：一是总体的实体性，即总体是指研究对象物质实体的集合；另一是指标性，即统计分析所关注的是定义在物质实体上的可数量化的指标．

显然，在对这个数量化指标进行观测时，由于每个个体的出现是随机的，所以这个数量化指标是一个随机变量 X．总体概念的要旨是：**总体是一个随机变量**．因此，在后面的讨论中往往将总体称为(随机)变量．统计分析的根本目标就是通过对变量的观测，指出变量的概率分布及其数字特征．

总体不仅可以用随机变量表示，也可以用它们的分布函数 $F(x)$表示．有了这个观点，就可以在概率论的基础之上展开统计研究．两个总体即使其所含个体的性质根本不同，只要有统一的概率分布，则在统计学中就视为同类总体．

总体类型即变量的概率分布类型，常见的有：正态分布，指数分布，均匀分布，β 分布，γ 分布，对数正态分布，瑞利分布，威布尔分布等连续型分布；还有二项分布，泊松分布，几何分布，超几何分布，离散均匀分布，负二项分布等离散型分布．为后续讨论方便，下面列出这些常见分布类型的数学定义(概率密度函数)及其数学期望与方差，见

表 2.1.

表 2.1　常见概率分布类型的数学定义、数学期望与方差

分布类型	分布的数学定义(密度函数)	数学期望	方　差
正态分布	$f(x\mid\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty<x<+\infty,\ -\infty<\mu<+\infty,\ \sigma>0$	μ	σ^2
指数分布	$f(x\mid\mu)=\frac{1}{\theta}e^{-\frac{x}{\theta}}I_{x\geqslant 0}(x)$ $x\geqslant 0,\ \theta\in N^+$	θ	θ^2
均匀分布	$f(x\mid a,b)=\frac{1}{b-a}I_{[a,b]}(x)$ $a\leqslant x\leqslant b,\ a<b$	$\frac{a+b}{2}$	$\frac{b-a}{12}$
β 分布	$f(x\mid a,b)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}I_{[0,1]}(x)$ $0\leqslant x\leqslant 1,\ a,b>0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$
γ 分布	$f(x\mid a,b)=\frac{1}{b^a\Gamma(a)}x^{a-1}e^{-\frac{x}{b}}I_{x\geqslant 0}(x)$ $x\geqslant 0,\ a,b>0$	ab	ab^2
对数正态分布	$f(x\mid\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$ $x>0,\mu>0,\sigma>0$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2\mu+2\sigma^2}-e^{2\mu+\sigma^2}$
瑞利分布	$f(x\mid b)=\frac{x}{b^2}e^{-\frac{1}{2}\left(\frac{x}{b}\right)^2}$ $-\infty<x<+\infty,b\neq 0$	$b\left(\frac{\pi}{2}\right)^{\frac{1}{2}}$	$\frac{2-\pi}{2}\cdot b^2$
威布尔分布	$f(x\mid a,b)=abx^{b-1}e^{-ax^b}I_{x>0}(x)$	$a^{-1/b}\Gamma(1-b^{-1})$	$a^{-2/b}[\Gamma(1+2b^{-1})-\Gamma^2(1+b^{-1})]$
二项分布	$f(x\mid n,p)=\binom{n}{x}p^x(1-p)^{n-x}I_{(0,1,2,\cdots,n)}(x)$ $x=0,1,\cdots,n,n\in N^+,\ 0\leqslant p\leqslant 1$	np	$np(1-p)$
泊松分布	$f(x\mid\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}I_{(0,1,2,\cdots)}(x)$ $x=0,1,2,\cdots,\lambda>0$	λ	λ
几何分布	$f(x\mid p)=p(1-p)^xI_{(0,1,2,\cdots)}(x)$ $0\leqslant p\leqslant 1$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

续表 2.1

分布类型	分布的数学定义(密度函数)	数学期望	方差
超几何分布	$f(x\|M,K,n)=\dfrac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}I_{(0,1,\cdots,r)}(x)$ $M,K,n\in N^+,n\leqslant M,r=\min(K,n)$	$n\cdot\dfrac{K}{M}$	$n\cdot\dfrac{K}{M}\cdot\dfrac{M-K}{M}\cdot\dfrac{M-n}{M-1}$
离散均匀分布	$f(x\|N)=\dfrac{1}{N}I_{(1,2,\cdots,N)}(x)$ $N\in N^+$	$\dfrac{N+1}{2}$	$\dfrac{N^2-1}{12}$
负二项分布	$f(x\|r,p)=\binom{r+x-1}{x}p^r(1-p)^xI_{(0,1,\cdots)}(x)$ $0\leqslant p\leqslant 1$	$\dfrac{r(1-p)}{p}$	$\dfrac{r(1-p)}{p^2}$

关于概率分布及其数字特征的详细讨论属于概率论的范畴，这里仅就 MATLAB 对上述内容的描述作简单介绍.

MATLAB 为常见的概率分布提供了下列 5 类函数：

① 概率密度函数(pdf). 求随机变量 X 在 x 点处的概率密度值 $y=p(x)$.

② 累积分布函数(cdf). 求随机变量 X 在 x 点处的分布函数值

$$F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}p(u)\mathrm{d}u.$$

② 逆累积分布函数(inv). 求随机变量 X 在 x 点处的分布函数的反函数值 $x=F^{-1}(f)$.

④ 均值与方差计算函数(stat). 求给定分布的随机变量 X 的数学期望 $E(X)$ 和方差 $\mathrm{var}(X)$.

⑤ 随机数生成函数(rnd). 模拟生成指定分布的样本数据.

具体函数的命名规则是：

函数名＝分布类型名称＋函数类型名称(pdf、cdf、inv、stat、rnd)

其中，分布类型名称如下：

分布类型	MATLAB 名称
正态分布	norm
指数分布	exp
均匀分布	unif
β分布	beta
γ分布	gam
对数正态分布	logn

瑞利分布	rayl
威布尔分布	weib
二项分布	bino
泊松分布	poiss
几何分布	geo
超几何分布	hyge
离散均匀分布	unid
负二项分布	nbin

例如，normpdf、normcdf、norminv、normstat 和 normrnd 分别是正态分布的概率密度、累积分布、逆累积分布、数字特征和随机数生成函数.

关于这5类函数的语法，请详见本书附录B，或参见文献[4]. 快捷的学习可借助 MATLAB 的系统帮助，通过指令 doc 获得具体函数的详细信息，语法是：

doc <函数名>

关于本书中涉及的统计分析指令的深入学习均可按此提示进行，后文不再赘述.

正态分布在统计分析中占有中心地位. 下面对正态分布的性质进行直观回顾.

【例 2.1】 绘制正态分布的密度函数、分布函数曲线，并求均值与方差.

```
clear
mu = 2.5; sigma = 0.6; % 设定正态分布的分布参数 μ 和 σ
x = (mu - 4 * sigma): 0.005: (mu + 4 * sigma); % 设定绘图区域 μ±4σ
y = normpdf(x,mu,sigma); % 计算与 x 对应的概率密度值
f = normcdf(x,mu,sigma); % 计算与 x 对应的分布函数值
plot(x,y,'-g',x,f,':b') % 绘制正态分布的密度函数、分布函数曲线
[M,V] = normstat(mu,sigma) % 求数学期望与方差的值
legend('pdf','cdf', -1) % 添加图例
```

上述指令的运行结果见图 2.1 及：

```
M =
    2.5000
V =
    0.3600
```

从图 2.1 中可以看出，正态密度曲线是关于 $x=\mu$ 对称的钟形曲线(两侧在 $\mu\pm\sigma$ 处各有一个拐点)，正态累积分布曲线当 $x=\mu$ 时 $F(x)=0.5$.

【例 2.2】 正态分布参数对密度曲线的影响(绘图指令集 M-脚本文件 normplot_1 见本书附录 C).

从图 2.2 中可以看出，μ 决定了图形的中心位置，σ 决定了图形中峰的陡峭程度，

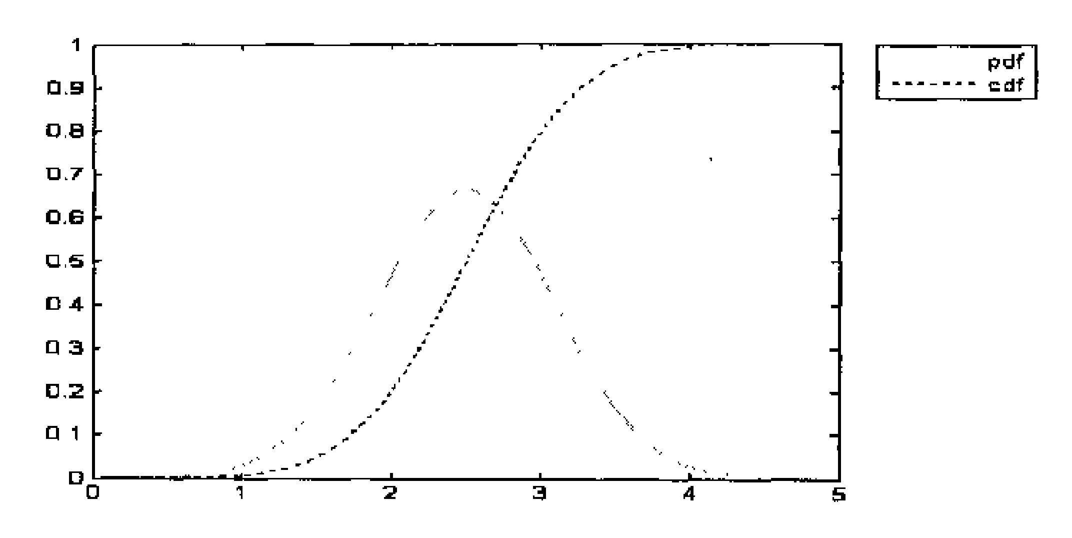


图 2.1　正态分布的密度函数与分布函数曲线

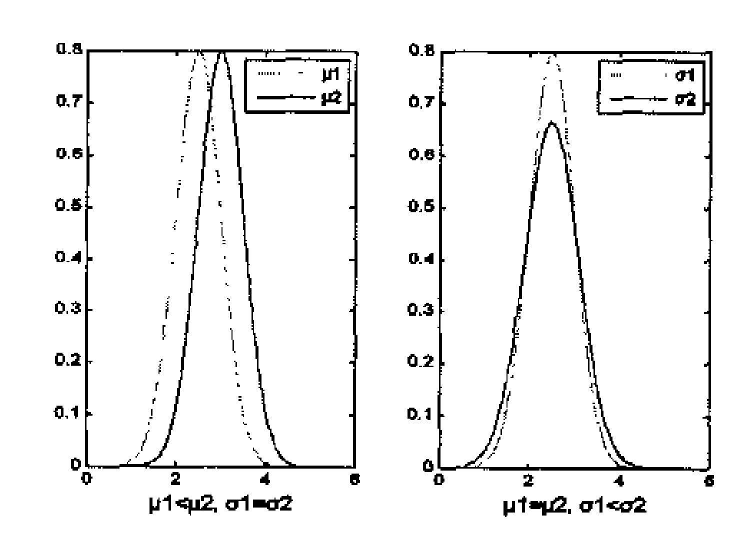


图 2.2　正态分布参数对密度曲线的影响

当 σ 较大时，图形趋于平缓；当 σ 较小时，图形趋于陡峭．也就是说，μ 决定了分布的中心位置，σ 反映了分布的分散或集中程度．

【例 2.3】 正态分布参数 μ 和 σ 对变量 X 取值规律的约束——3σ 准则(绘图指令集 M-脚本文件 normplot_2 见本书附录 C)．

从图 2.3 中可以看出，正态分布在均值 μ 处密度最大，即正态随机变量 X 最有可能在点 μ 附近取值；在 μ 两侧，$\pm\sigma$ 的范围内取值概率为 0.6826，$\pm 2\sigma$ 的范围内取值概率为 0.9544，$\pm 3\sigma$ 的范围内取值概率为 0.9974．虽然正态随机变量 X 可能在整个数轴上取值，但是其取值几乎全部集中在区间 $[\mu-3\sigma,\ \mu+3\sigma]$ 内，统计学称之为“3σ 准则”．

【例 2.4】 3σ 准则的应用．

已知测量值 $Y \sim N(0.2,\ 0.052)$，今发现十次测量中有一个数据是 0.367，问是否

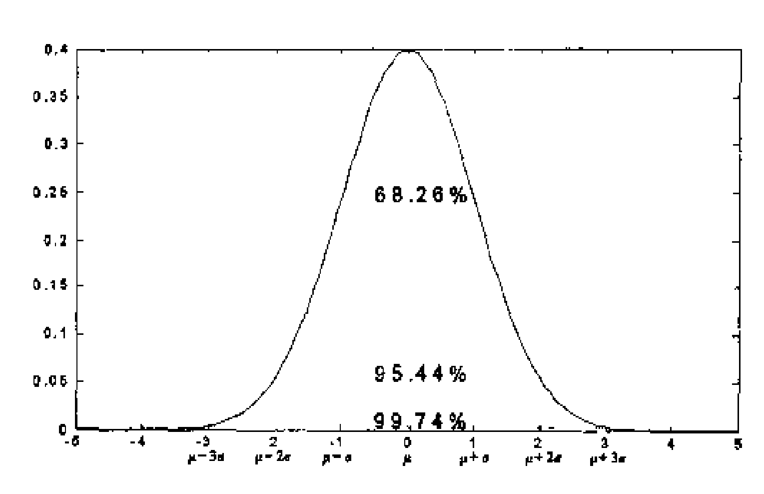


图 2.3　正态分布参数对密度曲线的影响

可认为异常而予以剔除?

解　由 3σ 准则知道，测量值以超过 0.997 的概率在 $0.2\pm0.05\times3$ 之间，即在 0.05 与 0.35 之间. 由于 $0.367>0.35$，故应剔除这个数据.

在自然界和社会领域常见的变量中，很多都属于用正态分布刻画的范围. 例如，人的身高高低不等，但中等身材的占大多数，特高或特矮的只是少部分，而且较高和较矮的人数大致相近. 又如，一个班的考试成绩，很好和很差的人数都不多，多数处在中间状态，但以一个平均分数为中心去观察，高于它和低于它的分布情况相似，等等. 进一步地，中心极限定理的研究表明，一个变量如果受到大量微小、独立的随机因素的影响，或者说，一个随机变量可以表示为若干个独立随机变量之和，那么，这个变量一般近似为一个正态变量.

2.1.2　变量的观测与数据

总体是一个随机变量(以下简称为变量)，这是统计学的一个基石性的概念. 人们对变量的认识是通过对变量的观测实现的，这一过程称为抽样.

抽样是指为获得有关变量的信息，按一定的规则对变量进行的观察和试验. 抽样的结果称为样本. 样本规定了如何对变量 X 进行 n 次观测以获得关于这个变量相关信息资料，n 称为样本容量.

理解样本的概念，关键是理解样本的两阶段性，即需要理清理论上的样本和样本数据的联系与区别.

① **理论上的样本是一个 n 维随机向量**$(X_1, X_2, \cdots, X_n)$. 人们对变量 X 的认识是通过多次观测实现的. 对变量 X 的第 i 次观测记为 X_i，$i=1,2,\cdots,n$. 容易理解，在进行具体的观测之前，X_i 也是一个随机变量. 因为在具体观测时，可以对这个个体进行观测，也可以对另一个个体进行观测，这是具有随机性的，也就是说 X_i 的取值具有随机

性. 将各次观测表示为一个向量$(X_1, X_2, \cdots, X_n)$, 是因为在对变量 X 的认识过程中需要它们发挥整体作用.

② **样本数据是样本的一次观察值**. 如果完成了对样本$(X_1, X_2, \cdots, X_n)$的一次具体观测, 就得到 n 个具体的观测或试验数据$(x_1, x_2, \cdots, x_n)$, 称为**样本的一次观察值或样本数据**, 简称为数据.

样本是一个抽象的理论概念, 是联系数据资料与变量特征的桥梁.

总体(随机变量) —抽样的设计→ 样本(n 维随机向量) —抽样的实施→ 数据(n 维数值向量)

数据 → 总体: 统计描述与推理(实现从个别到整体的认识)

在实际应用中, 抽样是一个复杂的过程, 需要进行精心设计和严谨的组织实施. 在数理统计中一般不对抽样问题进行过多的讨论, 只是假定抽样满足如下的基本要求: **样本的各个分量 X_1, X_2, $\cdots$, X_n 相互独立且与变量 X 同分布**. 满足这一要求的抽样称为简单随机抽样, 由简单随机抽样抽取的样本称为简单随机样本. 在数理统计的讨论中, 若不特别说明, "样本"一词均指简单随机样本.

样本的概率分布是统计分析基本的理论依据. 由样本所满足的基本条件易知, 若变量 X 的分布函数为 $F(x)$(概率函数或概率分布律为 $p(x)$), 则称 $X_1, X_2, \cdots, X_n$ 为来自总体 $F(x)$(或 $p(x)$)的样本, 记为 $X_1, X_2, \cdots, X_n \text{ i.i.d} \sim F(x)$(或 $p(x)$), 并且样本的联合分布函数(或联合概率函数)为

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i) \quad (\text{或 } p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i)).$$

由此可见, 变量的概率分布决定了样本的概率分布, 进而也就决定了样本数据的统计规律, 也就是样本取到样本数据的规律. 统计学的基本思想就是通过样本数据统计规律的这种规定性来达到认识变量的目的(反向应用). 也就是说, 通过对变量的观测以获取相关的数据信息, 利用变量、样本和数据之间的内在联系, 由样本数据去推断变量的特征与变化规律. 简要地说, **数据是认识变量的基本依据**.

统计分析的方法往往受变量测度性质的制约, 而变量的类型又决定数据的性质. 因此, 统计分析要谨慎选择与变量类型和数据性质相适应的方法.

依变量测度性质的不同, 变量可以区分为如下三种类型.

(1) 定性变量(分类变量)

对变量进行观测时仅可作类属的判定. 如学生的性别.

(2) 顺序变量

对变量进行观测时仅可作顺序的比较. 如按等级评定的学生考查成绩.

(3) 定量变量

对变量进行观测时存在一种可数量化的尺度，在这个尺度下可以确定一个观测值与另一个观测值数量上的差异或比率关系. 定量变量依观测是否存在绝对零点可以进一步区分为等距变量(无绝对零点)和比率变量(有绝对零点). 如学生考试的百分制成绩(等距变量)，植物的高度(比率变量)，员工的工资金额(比率变量)等.

进而，数据可以按其所属变量的不同而分为定性数据、顺序数据和定量数据.

(1) 定性数据

定性数据是事物类属性(非数值性)的描述，从属于定性变量. 如对学生性别变量进行观测，男性赋值为“1”，女性赋值为“0”，则数据 0 和 1 就是分类数据. 分类数据仅可以按其所属类别进行计数(统计频数)，而对计数的结果可以进行加法(合计)或百分数运算.

(2) 顺序数据

顺序数据是事物优劣属性(非数值性)的描述，从属于顺序变量. 如按等级评定的学生考查成绩，对某门课程优秀赋值为“5”，良好赋值为“4”等，则数据 1，2，3，4，5 就是顺序数据. 这类数据通常是按照一定的准则测量出来的，但是测量准则既无绝对零点又无相等的尺度单位. 如果甲学生的成绩被评为“5”，乙学生的成绩被评为“4”，我们只能说甲的成绩比乙的成绩好，但说不出甲的成绩比乙的成绩好多少或好多少倍. 也就是说，这类数据只能进行计数和大小的比较，不能进行加、减、乘、除运算.

(3) 定量数据

定量数据是事物量的属性(数值性)的描述，从属于定量变量. 包括如下两种类型.

① 等距数据，从属于等距变量的观测数据. 如按百分制评定的学生考试成绩就是一个等距变量. 人们通常认为试卷中的每一分值所表征的、对学生在该试卷范围内的知识和能力的要求是一样的，因此由试卷测量出的成绩数据有相等的尺度单位. 如果甲学生得 80 分，乙学生得 40 分，比较时我们不仅能说在这次考试中甲的成绩比乙的成绩好，还能说出甲的分数比乙的分数多 40 分，但是我们不能说甲的知识和能力水平是乙的 2 倍. 即使某个学生在考试中得了 0 分，也不能说他在该课程中没有一点儿知识和能力，因为这类测量数据不是从绝对零点计算起的，它仅在某个区间(如该份试卷所考查的知识范围)内具有相等的尺度单位，但我们不能确定这个尺度单位与区间内外统一观察时可能采用的尺度单位之间的比例关系. 概括地说，等距数据是在没有绝对零点但有相等尺度单位的测量过程中得到的，可以进行计数、大小的比较和加、减运算，但不能进行乘、除运算.

② 比率数据，从属于比率变量的观测数据. 一般认为，由物理方法测量得到的数据是比率数据，如物体的质量、生物的寿命、生活中某种消费品的消费量等. 由于这类数据存在统一的、物理的度量尺度，有绝对零点和相同的尺度单位，可以得到加、减、乘、除运算结果有意义的解释，因此，比率数据可以进行计数和大小的比较，以及加、减、乘、除运算.

在数据处理中还应当注意数据的连续性问题. 通常, 物理方法的测量尺度单位往往能够无限分割成更细小的单位, 因此比率数据是连续性数据. 等距数据也可以作为连续性数据进行处理. 但是, 定性变量下的计数数据则是离散性数据.

2.2 统计分析的基本工具

2.2.1 统计量

统计量是统计分析的基本工具.

统计量是指样本的不含其他未知参数的函数. 统计量概念的要点是"不含其他未知参数", 即只要给定样本数据, 则统计量的函数值(统计量的观测值)就能够唯一地确定下来.

统计分析技术在一定程度上可以说是统计量的构造技术. 学习过程中要高度重视针对某种特定的问题是如何构造相关统计量的. 本小节仅讨论几类基本的统计量, 这是在特定的问题中构造相关统计量的基础材料.

(1) 样本矩

样本矩是最基本、常用的一类统计量, 主要包括如下两种.

① 样本 k 阶(原点)矩. 设 X_1, X_2, …, X_n i.i.d～X, 则称

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad (k = 1,2,\cdots)$$

为变量 X 的样本 k 阶(原点)矩. A_k 的观测值记为 μ_k.

特别地, 样本的 1 阶矩

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

称为样本均值, 它是最重要的统计量之一, 反映了变量 X 取值集中程度的信息. $\overline{X}$ 的观测值用 $\bar{x}$ 表示.

② 样本 k 阶中心矩. 设 $X_1, X_2, \cdots, X_n$ i.i.d～X, 则称

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \quad (k = 1,2,\cdots)$$

为变量 X 的样本 k 阶中心矩. B_k 的观测值记为 ν_k.

特别地, 样本的 2 阶中心矩

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

称为样本方差, 它也是最重要的统计量之一, 反映了变量 X 取值分散程度的信息. S^2

的观测值用 s^2 表示.

值得注意的是, 在实际应用中, 常用样本的修正方差

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

替代 S^2(以下若无特别说明, "样本方差"一词均指样本的修正方差, 仍记为 S^2). 样本方差的算术根称为样本标准差, 记为 S, 即 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

(2) 顺序统计量

顺序统计量是另一类最基本、常用的统计量.

设 $X_1, X_2, \cdots, X_n$ i.i.d $\sim X$, $(x_1, x_2, \cdots, x_n)$是$(X_1, X_2, \cdots, X_n)$的任意一次观测值. 记 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$是 $x_1, x_2, \cdots, x_n$ 的一个排列, 并且 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. 若令 n 维随机向量$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$总是以$(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$为观测值, 则称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$为变量 X 的一个顺序统计量.

由顺序统计量出发, 可以构造许多有用统计量, 例如:

① 样本最大值 $X_{\max} = \max(X_1, X_2, \cdots, X_n) = X_{(n)}$;

② 样本最小值 $X_{\min} = \min(X_1, X_2, \cdots, X_n) = X_{(1)}$;

③ 样本极差 $R = \max(X_1, X_2, \cdots, X_n) - \min(X_1, X_2, \cdots, X_n) = X_{(n)} - X_{(1)}$;

④ 样本中位数 $\hat{m} = \begin{cases} X_{\frac{n+1}{2}}, & n \text{ 为奇数}, \\ \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right), & n \text{ 为偶数}. \end{cases}$

2.2.2　数据特征的度量及其 MATLAB 函数

统计量最基本的应用就是对数据特征的度量. MATLAB 定制了样本数据的一些常用度量性的统计描述函数, 下面就最常用的部分分别予以介绍.

(1) 数据集中性的度量

数据集中性的度量见表 2.2.

表 2.2　数据集中性的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本均值	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	mean
样本中值	$m_{0.5}$(参见样本的经验分位数)	median
样本几何均值	$\bar{x}_g = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$	geomean
样本调和均值	$\bar{x}_h = n\left(\sum_{i=1}^{n} x_i^{-1}\right)^{-1}$	harmmean

(2) 数据变异性的度量

数据变异性的度量见表 2.3.

表 2.3 数据变异性的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本方差	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	var
样本标准差	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	std
样本极差	$R = x_{(n)} - x_{(1)}$	range
样本内四分位数间距	$I = m_{0.75} - m_{0.25}$(参见样本的经验分位数)	iqr

(3) 数据分布特征的度量

数据分布特征的度量见表 2.4.

表 2.4 数据分布特征的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本的经验分位数	$m_p = \begin{cases} x_{([np+1])}, & np \notin N, \\ 0.5(x_{(np)} + x_{(np+1)}), & np \in N \end{cases}$	prctile
样本峰度	$KU = \frac{B_4}{B_2^2}$	kurtosis
样本偏度	$SK = \frac{B_3}{B_2^{\frac{3}{2}}}$	skewness

(4) 两组数据线性相依程度的度量

两组数据线性相依程度的度量见表 2.5.

表 2.5 两组数据线性相依程度的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本协方差	$c = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(y_j - \bar{y})$	cov
样本相关系数	$r = \frac{c}{s_x s_y}$	corrcoef

2.3 统计分析的推理基础

2.3.1 常用的统计分布与 α 分位数

统计量作为随机变量的函数，也是随机变量，自然要服从某种概率分布. 统计量的概率分布称为抽样分布. 统计推理的品质完全取决于其所依赖的抽样分布的性质. 通

常，抽样分布区分为精确分布(当总体 X 的分布类型已知时，对任一自然数 n，都能导出统计量的分布的明显表达式)和渐近分布(借助于中心极限定理)两类. 依据精确分布可进行小样本统计推理，而渐近分布则是大样本统计推理的理论基础.

本小节介绍抽样分布的概率基础——常用的统计分布.

2.3.1.1　χ^2 分布

定义 2.1　如果随机变量 X 的概率密度为

$$f(x,n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, & x\geqslant 0,\\ 0, & x<0,\end{cases}$$

其中，参数 n 取正整数，$\Gamma(x)$ 是通过积分 $\Gamma(x)=\int_0^{\infty}\mathrm{e}^{-t}t^{x-1}\mathrm{d}t\,(x>0)$ 定义的伽玛函数，则称 X 服从参数为 n 的 χ^2 分布，记为 $X\sim\chi^2(n)$，并称参数 n 为 χ^2 分布的自由度，$f(x)$ 称为 χ^2 分布的密度函数.

χ^2 分布的密度函数曲线见图 2.4(绘图指令集 M-脚本文件 chi2plot 见本书附录 C).

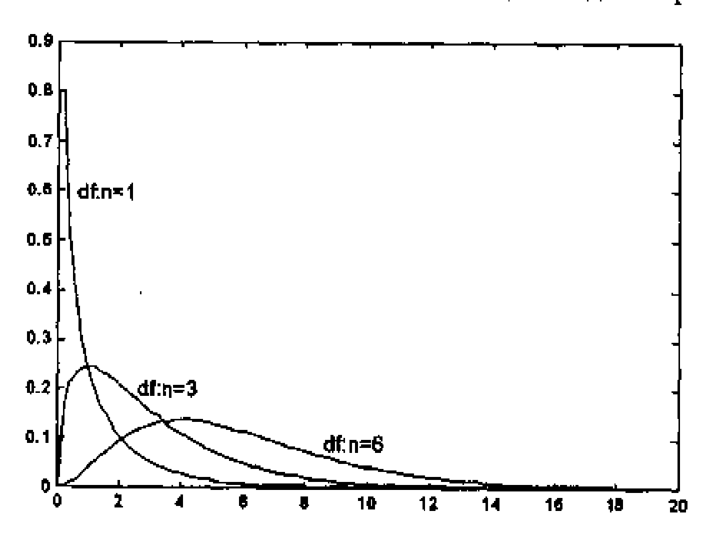


图 2.4　χ^2 分布的密度函数曲线

不难求出 χ^2 分布的数学期望和方差：设 $X\sim\chi^2(n)$，则 $E(X)=n$，$\mathrm{Var}(X)=2n$.

下面不加证明地列出关于 χ^2 分布的两个常用定理.

定理 2.1 (χ^2 分布的可加性)　设 $X_1\sim\chi^2(n_1)$，$X_2\sim\chi^2(n_2)$，且 X_1, X_2 相互独立，则 $X_1+X_2\sim\chi^2(n_1+n_2)$.

定理 2.2 (χ^2 分布的统计生成定理)　设 $X_1, X_2, \cdots, X_n$ i.i.d $\sim N(0,1)$，令 $\kappa=X_1^2+X_2^2+\cdots+X_n^2$，则统计量 $\kappa\sim\chi^2(n)$.

推论 设 $X_1, X_2, \cdots X_n$ i.i.d~$N(\mu, \sigma^2)$, 令 $\kappa = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$, 则 $\kappa \sim \chi^2(n)$.

2.3.1.2 t 分布

定义 2.2 如果随机变量 X 的概率密度为

$$f(x, n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{n\pi}}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

则称 X 服从参数为 n 的 t 分布，记为 $X \sim t(n)$，并称参数 n 为 t 分布的自由度，$f(x)$ 称为 t 分布的密度函数.

t 分布的密度函数曲线见图 2.5(绘图指令集 M-脚本文件 tplot 见本书附录 C).

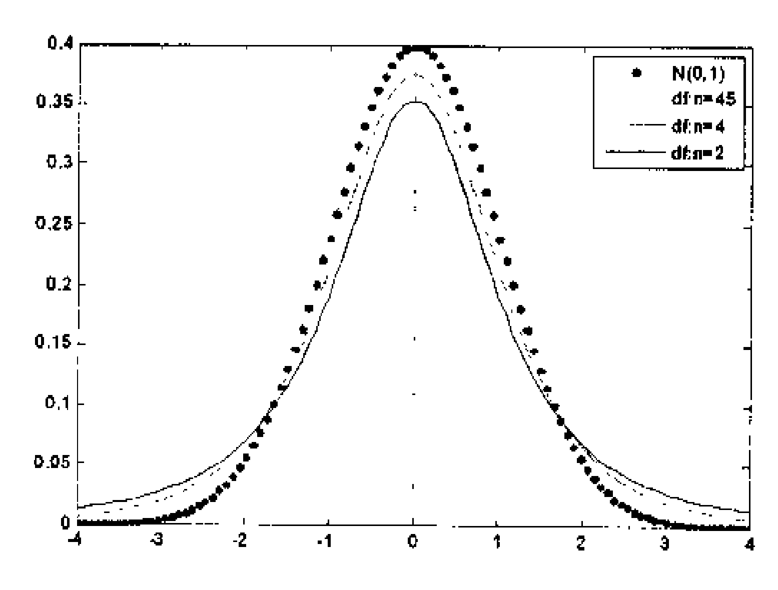


图 2.5 t 分布的密度函数曲线

由图 2.5 可以看出，t 分布的密度函数具有对称性和渐近正态性.

① 对称性. t 分布的密度函数 $f(x)$ 关于 $x = 0$ 对称，且 $\lim_{|x| \to \infty} f(x) = 0$.

② 渐近正态性. 当 n 充分大(≥45)时，t 分布近似于标准正态分布，即

$$\lim_{n \to \infty} f(x, n) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

但对于较小的 n，t 分布与标准正态分布相差很大.

不难求出 t 分布的数学期望和方差：设 $T \sim t(n)$，则当 $n > 2$ 时，有 $E(T) = 0$, $\mathrm{Var}(T) = \frac{n}{n-1}$.

下面不加证明地列出关于 t 分布的一个常用定理.

定理 2.3 (t 分布的统计生成定理) 设统计计量 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X 与 Y

相互独立，则统计量 $T=\frac{X}{\sqrt{Y/n}}\sim t(n)$.

2.3.1.3　F 分布

定义 2.3　如果随机变量 X 的概率密度为

$$f(x,n_1,n_2)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\cdot\left(\dfrac{n_1}{n_2}\right)\left(\dfrac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x\geqslant 0,\\ 0, & x<0,\end{cases}$$

则称 X 服从参数为 n_1，n_2 的 F 分布，记为 $F\sim F(n_1,n_2)$，并称参数 n_1，n_2 为 F 分布的第一、第二自由度，$f(x)$称为 F 分布的密度函数.

由定义 2.3 可知，F 分布具有倒数对称性：若 $F\sim F(n_1,n_2)$，则 $\frac{1}{F}=\frac{Y/n_2}{X/n_1}\sim F(n_2,n_1)$.

F 分布的密度函数曲线如下.

① F 分布的密度函数曲线形态(绘图指令集 M-脚本文件 fplot_1 见本书附录 C)见图 2.6.

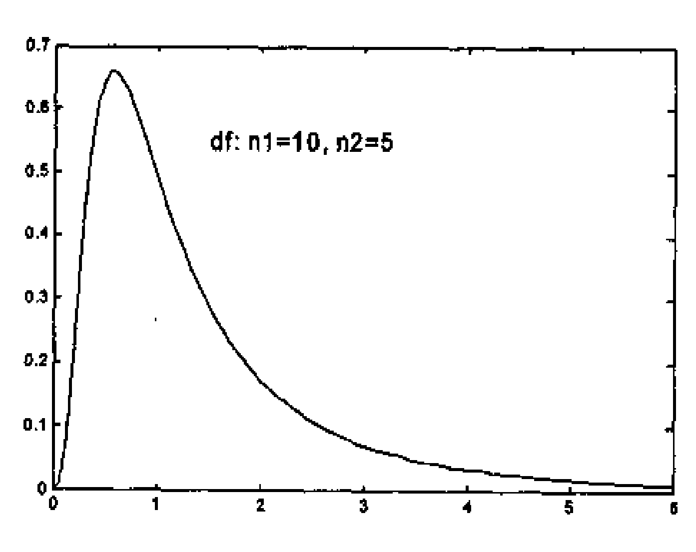


图 2.6　F 分布的密度函数曲线

② 自由度对 F 分布密度曲线形态的影响(绘图指令集 M-脚本文件 fplot_2 见本书附录 C)见图 2.7.

不难求出 F 分布的数学期望和方差：设 $F\sim F(n_1,n_2)$，则

$$E(F)=\frac{n_2}{n_2-2}\quad(n_2>2),\quad \mathrm{Var}(F)=\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}\quad(n_2>4).$$

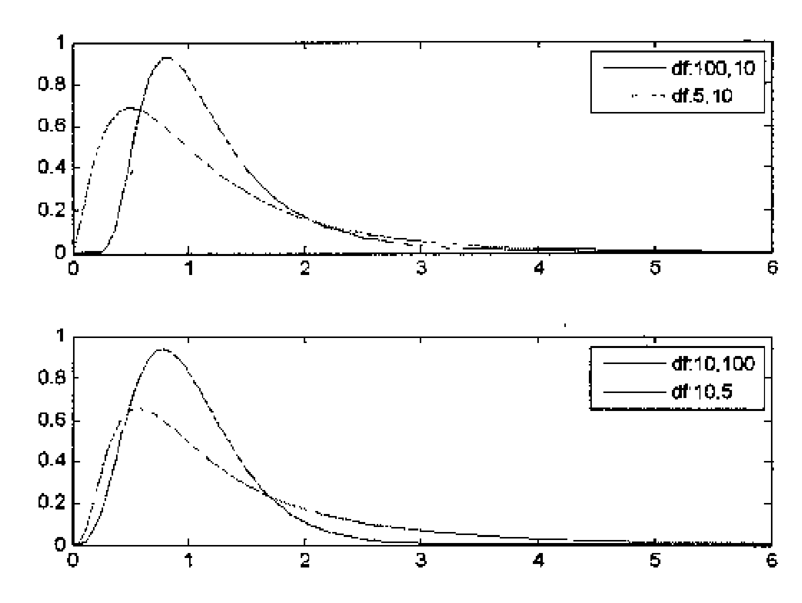


图 2.7 自由度对 F 分布密度曲线形态的影响

下面不加证明地列出关于 F 分布的一个常用定理.

定理 2.4 (F 分布的统计生成定理) 设统计量 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且 X 与 Y 相互独立, 则统计量 $F = \dfrac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$.

上述三种统计分布亦称为中心 χ^2 分布、t 分布和 F 分布. 需要指出的是, 由密度函数给出的定义与统计生成定理是等价的, 在许多教材中往往直接用统计生成定理作为定义使用. 对三个统计生成定理的证明感兴趣的读者请参见文献[1], 与之相应的有非中心 χ^2 分布、t 分布和 F 分布. 非中心统计分布与中心统计分布一样, 在统计推断中也发挥着重要的作用.

关于非中心分布, 这里仅给出它们的统计生成定义.

定义 2.4 (非中心 χ^2 分布) 设 $X_i \sim N(\mu_i, \sigma^2)$ $(i = 1, 2, \cdots, n)$, 且相互独立, 令 $\kappa = \dfrac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2$, 则称 κ 服从自由度为 n、非中心参数为 μ 的非中心 χ^2 分布, 记为 $\kappa \sim \chi^2(n, \mu)$, 其中 $\mu^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}\mu_i^2$.

定义 2.5 (非中心 t 分布) 设统计量 $X \sim N(\mu, 1)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 令 $T = \dfrac{X+\mu}{\sqrt{Y/n}}$, 则称 T 服从自由度为 n、非中心参数为 μ 的非中心 t 分布, 记为

$$T \sim t(n, \mu).$$

定义 2.6 (非中心 F 分布) 设统计量 $X \sim \chi^2(n_1, \mu)$, $Y \sim \chi^2(n_2)$, 且 X 与 Y 相互独立, 令 $F = \dfrac{X/n_1}{Y/n_2}$, 则称 F 服从自由度为(n_1, n_2)、非中心参数为 μ 的非中心 F 分布,

记为 $F \sim F(n_1, n_2, \mu)$.

MATLAB 为三大类统计分布也提供了 pdf、cdf、inv、stat 和 rnd 类函数，相应的分布类型名称如下：

分布类型	MATLAB 名称
χ^2 分布	chi2
t 分布	t
F 分布	f
非中心 χ^2 分布	ncx2
非中心 t 分布	nct
非中心 F 分布	ncf

2.3.1.4 统计量的渐近分布

在大多数场合，精确的抽样分布不易求出，或者求出来的精确分布过于复杂而难以应用，这时人们借助于极限工具，寻求在样本容量无限大时统计量的极限分布.

假如这种极限分布能求出，那么在样本容量 n 较大时，可用此极限分布当做抽样分布的一种近似，这种分布称为渐近分布.

关于渐近分布，下面的定理是大样本统计分析的一个基石性的结论.

定理 2.5（Lévy-Lindeberg 中心极限定理） 设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$，μ, σ^2 分别是变量 X 的均值和方差，且 $0<\sigma^2<+\infty$，则对于充分大的 n，近似地有 $\bar{X} \sim N(\mu, \sigma^2/n)$.

对定理的证明参见文献[3]. 更多的关于统计量渐近分布的结论将在具体应用的场合给出.

在应用中，若推知统计量精确地服从上述三种统计分布中的某一种，则可在小样本条件下进行统计推理. 否则，必须在大样本($n>30$)条件下依据统计量的渐近分布进行统计推理.

2.3.1.5 α 分位数

在利用统计量与统计分布进行统计推理时，离不开概率分布的 α 分位数的概念.

定义 2.7（α 分位数） 设 $0<\alpha<1$，对于随机变量 X.

① α 分位数：满足 $P\{X \leqslant x_\alpha\} = \alpha$ 的点 x_α.

② 上侧 α 分位数：满足 $P\{X > x_\alpha\} = \alpha$ 的点 x_α.

③ 双侧 α 分位数：满足 $P\{X \leqslant x_{\alpha/2}\} = \alpha/2$ 且 $P\{X > x'_{\alpha/2}\} = \alpha/2$ 的点 $x_{\alpha/2}$ 与 $x'_{\alpha/2}$.

【例 2.5】 α 分位数概念图示与相关的 MATLAB 计算.

图 2.8 是 α 分位数概念示意图（绘图指令集 M-脚本文件 alphaplot 见本书附录 C），

分布类型是标准正态分布，$\alpha=0.05$.

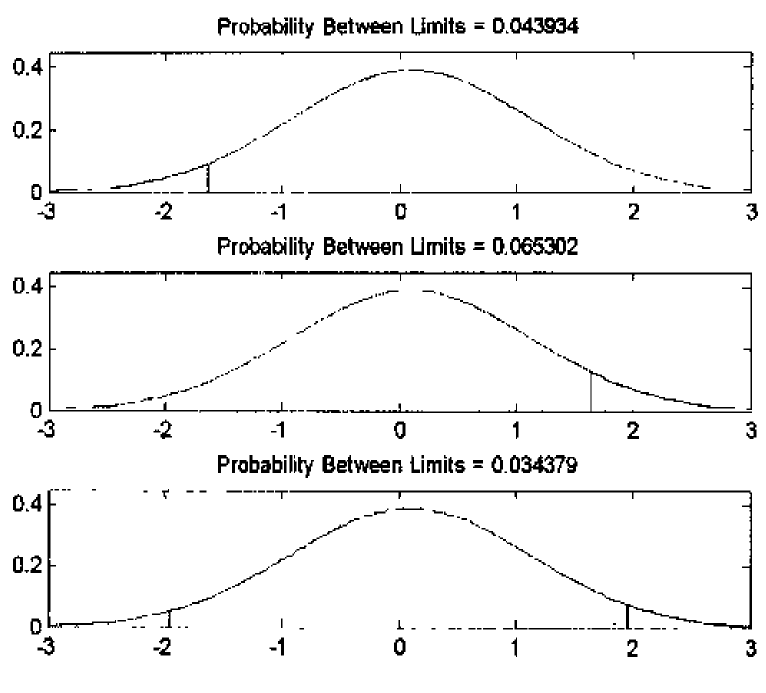


图 2.8 α 分位数概念示意图

在绘制指令集中，关键的几个 MATLAB 函数如下.

① 计算标准正态分布的 0.05 分位数. 如上侧 $\alpha/2$ 分位数的计算指令是

```
zalpha = norminv (0.975,0,1)
```

下面几个数值是标准正态分布的 0.05 分位数：

下侧分位数：−1.6449；

上侧分位数：1.6449；

双侧分位数：−1.9600，1.9600.

② 生成样本数据. 如生成 300 个标准正态分布的计算指令是

```
data = normrnd (0,1,300,1)
```

③ 绘制工序能力图(绘制由分位数控制的密度曲线下的面积图，用阴影表示，并计算样本数据落入控制区域的概率，显示在标题位置上). 计算指令是：

```
capaplot(data, [zalpha,inf])
```

在通常的数理统计教程中，有关分位数的值是通过查表求得的. 需要注意以下几点.

① 在 χ^2 分布上侧分位数表中可查到 $n=45$，对于 $n>45$，可以使用正态近似，一个较好的近似公式是 $\chi_\alpha^2(n)\approx\frac{1}{2}(u_\alpha+\sqrt{2n-1})^2$，其中 u_α 是标准正态分布的上 α 分位数.

② 在 t 分布上侧分位数表中只可查到 $n=45$，对于 $n>45$，可以使用正态近似 $t_\alpha(n)\approx u_\alpha$.

③ 容易证明，F 分布 α 分位数具有性质 $F_{1-\alpha}(n_1,n_2)=\dfrac{1}{F_\alpha(n_2,n_1)}$，此式可用来求 F 分布表中未列出的一些上 α 分位数，例如 $F_{0.95}(12,9)=\dfrac{1}{F_{0.05}(9,12)}=\dfrac{1}{2.80}=0.357$.

在统计应用中，由 MATLAB 进行数据处理和辅助分析，能够极大地简化基于查表计算某些冗繁的公式推导和变换.

2.3.2 基于正态分布的常用抽样分布

下面给出几个在小样本统计推断中常用的抽样分布定理.

定理 2.6（正态变量样本均值的抽样分布定理） 设 $X_1,X_2,\cdots,X_n$ i.i.d. $\sim N(\mu,\sigma^2)$，则

$$\overline{X}\sim N(\mu,\sigma^2/n),\quad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1).$$

推论 设 $X_1,X_2,\cdots,X_n$ i.i.d $\sim N(\mu,\sigma^2)$，则

$$\sum_{i=1}^{n}k_iX_i\sim N\Big(\mu\sum_{i=1}^{n}k_i,\sigma^2\sum_{i=1}^{n}k_i^2\Big).$$

定理 2.6 及其推论是正态分布的可加性在统计中的推广，这里不再赘述其证明.

定理 2.7（正态变量样本方差的抽样分布定理） 设 $X_1,X_2,\cdots,X_n$ i.i.d. $\sim N(\mu,\sigma^2)$，$\overline{X}$ 和 S^2 分别为样本均值和样本方差，则

① 样本均值 $\overline{X}$ 与样本方差 S^2 独立；

② $\dfrac{(n-1)S^2}{\sigma^2}=\dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$；

③ $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$.

统计分析的技术在一定程度上是统计量的构造技术. 在这里，以定理 2.7③中的统计量为例进行说明.

【例 2.6】 统计量构造技术示例一（兼做定理 2.7③的证明）.

在应用问题中，有时需要对变量的均值 μ 作出某种判断，这一问题往往转化为样本均值 $\overline{X}$ 和 μ 的比较（数学上，量值之间的比较通常是对比较对象之间的值差或比值的考查），并需要对比较的结果作出概率的判断（所谓概率的判断，是说判断在一定的概率意义下可能是正确的，或者说据此判断作出决策可能要承担风险），这就需要构造一个作为分析工具的统计量，并且这个统计量的概率分布是可知的（否则这个统计量在分析中

一无用处).

如果 $X_1, X_2, \cdots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$, 并且总体的方差 σ^2 已知, 则由定理 2.6 可选统计量为

$$U=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1).$$

如果总体的方差 σ^2 未知, 则使用统计量 U 将带来很大的麻烦, 因为在确定 $\overline{X}$ 和 μ 的关系时, 未知的 σ^2 是一个障碍. 此时, 容易想到的是用样本方差 S^2 替代 σ^2, 这又产生了新的问题: 如此替换后的统计量还能服从 $N(0,1)$分布吗?

然而, 我们知道 $\kappa=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$, 且 U 与 κ 二者独立, 于是由 t 分布的统计生成定理可知

$$T=\frac{U}{\sqrt{\kappa/(n-1)}}\sim t(n-1),$$

化简即得

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1).$$

在这里我们用到了定理 2.7 的结论①和②作为推理的依据. 关于这两个结论的证明涉及更多的基础知识, 稍复杂一些, 感兴趣的读者请参见文献[2].

定理 2.8 (两个正态变量样本均值差的抽样分布定理) 设总体 $X\sim N(\mu_1,\sigma^2)$, $Y\sim N(\mu_2,\sigma^2)$, 且 X 与 Y 相互独立, X_1, X_2, X_{n_1} i.i.d. $\sim N(\mu_1,\sigma^2)$, $Y_1, Y_2, \cdots, Y_{n_2}$ i.i.d $\sim N(\mu_2,\sigma^2)$, 则

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2).$$

其中

$$S_w=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}},\ S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2,\ S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2.$$

特别地, 当 $n_1=n_2=n$ 时, 上式成为

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2+S_2^2}{n}}}\sim t(2n-2).$$

【例 2.7】 统计量构造技术示例二(兼做定理 2.8 的证明).

在某些应用场合, 我们需要对两个变量的均值 μ_1 和 μ_2 进行比较并作出相应的概率

判断，这就需要构造一个作为分析工具的统计量，并且这个统计量的概率分布是可知的. 容易想到，变量均值 μ_1 和 μ_2 的亲疏程度可以用样本均值 $\bar{X}$ 和 $\bar{Y}$ 的亲疏程度近似描述，而

$$\bar{X} \sim N(\mu_1, \sigma^2/n_1), \quad \bar{Y} \sim N(\mu_2, \sigma^2/n_2),$$

所以

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2,\ \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

标准化后，有

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

如果两个变量的方差 σ_1^2，σ_2^2 已知并且相等(记为 σ^2)，则 U 就可以作为进一步分析所需要的统计量. 这里我们假定 σ_1^2，σ_2^2 未知，则 U 不可用. 考虑用 S_1^2，S_2^2 替代 σ_1^2，σ_2^2. 因为

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1), \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1),$$

由 $\chi^2(n)$ 的可加性知

$$\kappa = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_1 + n_2 - 2),$$

根据 t 分布的统计生成定理，有

$$T = \frac{U}{\sqrt{\kappa/(n_1 + n_2 - 2)}} \sim t(n_1 + n_2 - 2),$$

化简即得

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

其中

$$S_w = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

定理 2.9 (两个正态变量样本方差比的抽样分布定理)　设总体 $X \sim N(\mu_1,\ \sigma_1^2)$，$Y \sim N(\mu_2,\ \sigma_2^2)$，且 X 与 Y 相互独立，$X_1, X_2, \cdots, X_{n_1}$ i.i.d. $\sim N(\mu_1, \sigma^2)$，$Y_1, Y_2, \cdots, Y_{n_2}$ i.i.d. $\sim N(\mu_2, \sigma^2)$，S_1^2，S_2^2 分别是这两个样本的样本方差，则

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,\ n_2-1).$$

【例 2.8】 统计量构造技术示例三(兼做定理 2.9 的证明).

在某些应用场合，我们需要对两个变量的方差 σ_1^2 和 σ_2^2 进行比较并作出相应的概率判断，这就需要构造一个作为分析工具的统计量，并且这个统计量的概率分布是可知的. 容易想到，变量方差 σ_1^2 和 σ_2^2 的亲疏程度可以用样本方差 S_1^2 和 S_2^2 的亲疏程度近似描述，而

$$\kappa_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \kappa_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1),$$

且两者相互独立，由 F 分布的统计生成定理，有

$$F = \frac{\kappa_1/(n_1-1)}{\kappa_2/(n_2-1)} \sim F(n_1-1,\ n_2-1),$$

化简即得

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,\ n_2-1).$$

需要指出的是，在这里的讨论中假定了变量的均值是未知的.

2.3.3 顺序统计量的抽样分布

关于顺序统计量的分布问题，在许多初等概率论的教材中都有详细的讨论，下面不加证明地给出相关的结论.

定理 2.10 (顺序统计量的概率分布定理) 设变量 X 的分布函数为 $F(x)$(密度函数为 $f(x)$)，X 的样本顺序统计量为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，则

① $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 的联合分布的密度函数为 $n!\,f(x_1)\cdots f(x_n),\ x_1<\cdots<x_n$；

② $X_{(n)}$ 的分布函数为 $[F(x)]^n$，密度函数为 $nf(x)[F(x)]^{n-1}$；

③ $X_{(1)}$ 的分布函数为 $1-[1-F(x)]^n$，密度函数为 $nf(x)[1-F(x)]^{n-1}$；

④ $X_{(k)}$ 的密度函数为 $\dfrac{n!}{(k-1)!\,(n-k)!}f(x)[F(x)]^{k-1}[1-F(x)]^{n-k}$；

⑤ $(X_{(k)},\ X_{(l)})$ 的联合密度函数为 $\dfrac{n!}{(k-1)!\,(l-k-1)!\,(n-l)!}f(x_k)f(x_l)\cdot [F(x_k)]^{k-1}[F(x_l)-F(x_k)]^{l-k-1}[1-F(x_l)]^{n-l}$，其中，$k<l,\ x_k<x_l$.

关于上述结论的证明及其由顺序统计量诱导出的统计量(如样本极差)的分布，请读者参见文献[2].

习题2

1. 设 $X\sim N(0,\ 1)$，利用 MATLAB 求以下概率：

(1) $P\{X<1\}$； (2) $P\{X>1.5\}$； (3) $P\{-1<X\leqslant 2\}$.

2. 设 $X\sim N(5,3^2)$，利用 MATLAB 求概率 $P\{2<X\leqslant 10\}$.

3. 绘制自由度分别为 2, 5, 8 的 χ^2 分布的密度函数曲线，并分别求其均值与方差. 观察参数对密度曲线的影响.

4. 若 $T\sim t(n)$，试证明 $T^2\sim F(1,\ n)$.

5. 利用 MATLAB 计算 5, 13, 17, 29, 80, 150 这一组数据的算术均值、调和均值和几何均值，并比较它们之间的大小.

6. 设一批零件的长度 X(单位：cm)服从 $N(20,\ 0.2^2)$，现从这批零件中任取一件，求 ϵ 使 $P\{|X-20|\leqslant\epsilon\}=0.95$.

7. 设变量 $X\sim N(\mu,\ \sigma^2)$，$X_1,\ X_2,\ \cdots,\ X_n$ 为 X 的样本，问样本容量 n 至少应取多大才能使 $P\left\{\frac{S^2}{\sigma^2}\leqslant 1.5\right\}\geqslant 0.95$.

8. 写出计算正态分布 $N(3,\ 5^2)$的 0.1 上侧、下侧、双侧分位数的 MATLAB 计算指令.

9. 设 $X_1,\ X_2,\ \cdots,\ X_9$ 是来自正态变量 $X\sim N(0,\ 2^2)$的简单随机样本，求系数 a，b，c，使

$$X=a(X_1+X_2)^2+b(X_3+X_4+X_5)^2+c(X_6+X_7+X_8+X_9)^2$$

服从 χ^2 分布，并求其自由度.

10. 设 $X_1,X_2,\cdots,X_9$ 是来自标准正态变量 X 的简单随机样本，且

$$Y_1=\frac{1}{6}(X_1+\cdots+X_6),\quad Y_2=\frac{1}{3}(X_7+X_8+X_9),$$

$$S^2=\frac{1}{2}\sum_{i=7}^{9}(X_i-Y_2)^2,\quad Z=\frac{\sqrt{2}(Y_1-Y_2)}{S}.$$

求证：统计量 Z 服从自由度为 3 的 t 分布.

第3章 统计估计

统计估计是统计推断的主要内容，包括两个方面的任务：① 变量的分布形态未知，根据样本数据对变量的分布形态作出推测(估计)；② 变量的分布形态已知，即已知其概率分布函数(或概率分布律，或概率密度函数)的数学表达式，但是某些参数(或数字特征)未知，根据样本数据对未知的参数(或未知参数的函数)作出估计.

本章介绍统计估计的基本方法，包括频率直方图、经验分布函数与 box 图等对变量分布形态进行初步估计的方法，以及参数的矩估计方法和极大似然估计方法、估计量性能的评价、估计误差的分析与控制问题(参数的区间估计).

3.1 变量分布形态的估计

3.1.1 频率分布表与频率直方图

频率分布表是一种对连续性变量的观测数据进行分组整理和初步分析的一种重要的统计数据表. 频率直方图是频率分布表的图形化. 通过频率分布表与频率直方图，可以实现对变量分布形态(概率密度曲线)的初步估计. 掌握频率分布表的编制与频率直方图的绘制方法是统计应用的一项基本技能.

下面举例说明频率分布表的编制方法和频率直方图的绘制.

【例3.1】 钢材中的含硅量 X 是影响材料性能的一项重要因素. 在炼钢生产过程中，由于各种随机因素的影响，各炉钢的含硅量 X 是有差异的. 对含硅量 X 概率分布的了解是有关钢材料性能分析的重要依据. 某炼钢厂 120 炉正常生产的 25MnSi 钢的含硅量(单位：%)如下：

0.86	0.83	0.77	0.81	0.81	0.80	0.79	0.82	0.82	0.81
0.82	0.78	0.80	0.81	0.87	0.81	0.77	0.78	0.77	0.78
0.77	0.71	0.95	0.78	0.81	0.79	0.80	0.77	0.76	0.82
0.84	0.79	0.90	0.82	0.79	0.82	0.79	0.86	0.81	0.78
0.82	0.78	0.73	0.84	0.81	0.81	0.83	0.89	0.78	0.86
0.78	0.84	0.84	0.75	0.81	0.81	0.74	0.78	0.76	0.80
0.75	0.79	0.85	0.78	0.74	0.71	0.88	0.82	0.76	0.85

0.81	0.79	0.77	0.81	0.81	0.87	0.83	0.65	0.64	0.78
0.80	0.80	0.77	0.84	0.75	0.83	0.90	0.80	0.85	0.81
0.82	0.84	0.85	0.84	0.82	0.85	0.84	0.82	0.85	0.84
0.81	0.77	0.82	0.83	0.82	0.74	0.73	0.75	0.77	0.78
0.87	0.77	0.80	0.75	0.82	0.78	0.78	0.82	0.78	0.78

下面介绍如何编制频率分布表，以及绘制频率直方图的 MATLAB 实现方法.

首先，将上述 120 个含硅量数据载入 MATLAB 系统的工作内存(如预先编写数据文件 hgl.mat，保存到读者自己的工作路径下，然后运行下面的两条指令)：

```
clear
load hgl
```

下面介绍频率分布表的编制方法，其基本步骤如下.

(1) 数据分组

① 确定数据组个数. 根据样本容量 n 确定分组数 k，推荐公式为 $k=1.87(n-1)^{2/5}$.

② 计算极差. 计算公式为 $R_n=\max(x_1,x_2,\cdots,x_n)-\min(x_1,x_2,\cdots,x_n)$.

③ 确定组距. 计算公式为 $d\approx R_n/k$，一般取 d 为数据的最小测量单位的整数倍.

④ 确定各组端点. 计算公式为 $a_k=a_0+dk$ $(k=0,1,\cdots,n)$，其中，$a_0<\min\{x\}$，$a_n>\max\{x\}$. a_0 的确定方法：一般地取 a_0 比数据的最小值小半个测量单位.

(2) 统计各组频数

各组频数就是数据落入各个小组中的个数，记为 n_i.

上述计算的 MATLAB 实现由两步完成：第一步，先确定组数的推荐公式，求出分组数 k；第二步，其他的计算极差、确定组距、确定各组端点和统计各组频数的工作均可由 MATLAB 系统函数 hist 完成. hist 的输入参数通常有两个，第一个是数据向量，第二个是小组个数；hist 的输出参数有两个，第一个返回各组的数据频数，第二个返回各个数据组的区间位置值(组中值).

```
k = ceil(1.87 * (length(hgl) - 1)^0.4);
[ni, ak] = hist(hgl, k);
```

(3) 计算频率

① 计算各组频率. 计算公式为 $f_i=n_i/n$. MATLAB 计算指令为

```
fi = ni/length(hgl);
```

② 计算各组累积频率. 计算公式为 $F_i=\sum_{j=1}^{i}f_j(i=1,2,\cdots,k)$. MATLAB 计算指

令为

```
mfi = cumsum(fi);
```

(4) 编制频率分布表

逐一运行上述 MATLAB 计算指令，再运行指令

```
stats = [[1:k]',ak',ni',fi',mfi']
```

就可得到 120 炉 25MnSi 钢的含硅量数据的频率分布表，稍加整理即得到表 3.1.

表 3.1　120 炉 25MnSi 钢的含硅量数据频率分布表

组 序	组中值	频 数	频 率	累积频率
1	0.6519	2	0.0167	0.0167
2	0.6758	0	0	0.0167
3	0.6996	2	0.0167	0.0333
4	0.7235	2	0.0167	0.0500
5	0.7473	8	0.0667	0.1167
6	0.7712	29	0.2417	0.3583
7	0.7950	15	0.1250	0.4833
8	0.8188	36	0.3000	0.7833
9	0.8427	15	0.1250	0.9083
10	0.8665	6	0.0500	0.9583
11	0.8904	4	0.0333	0.9917
12	0.9142	0	0	0.9917
13	0.9381	1	0.0083	1.0000

接下来介绍频率直方图和累积频率折线图及其绘制方法.

频率直方图是连续性变量频率分布的图形化，累积频率折线图是累积频率分布的图形化.

在频率直方图中，横轴表示观测变量的观测值，每一个**小矩形的水平边长 = 组距**；纵轴表示各组数据的频率，由于概率密度曲线下方的面积恒等于 1，因此为保证直方图中所有矩形条面积之和也等于 1，规定每个**小矩形的高度 = 该组数据的频率/组距**.

用 MATLAB 绘制直方图的指令是 hist 或 histfit，但是需要指出的是，为观察上的方便，这两个指令绘制出的图形纵轴的刻度是频数值.

hist 前面已经见过，当有输出参数时，它将完成各组频数的统计工作；若无输出参数，则直接绘制频率直方图.

```
hist(hgl) % 画直方图
h = findobj(gca,'Type','patch'); % 为修饰图形提取指定属性对象的图形句柄
```

h(图形句柄是对图形进行细致修饰的操作对象，感兴趣的读者可参见文献[5])

```
set(h,'FaceColor','y','EdgeColor','b') % 修饰，设置直方图线条颜色与填充色
```

上述指令的运行结果见图3.1.

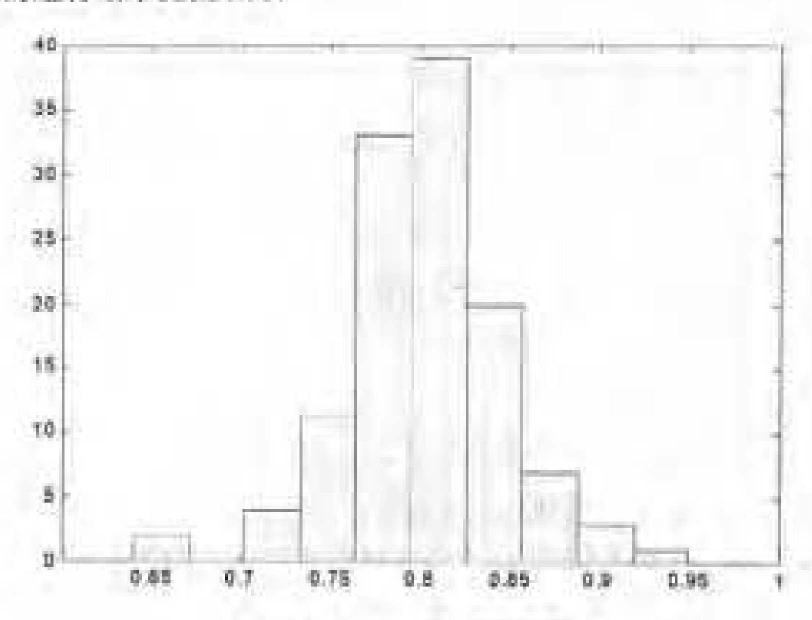


图3.1 hist指令绘制的直方图

histfit指令在绘制频率直方图的同时附加一条正态密度曲线，以供参考.

```
h = histfit(hgl, 13); % 画附正态参考曲线的直方图，并提取图形句柄h
set(h(1),'FaceColor','c','EdgeColor','w') % 修饰，设置直方图线条颜色与填充色
set(h(2),'Color','r') % 修饰，设置正态参考曲线的颜色
```

上述指令的运行结果见图3.2.

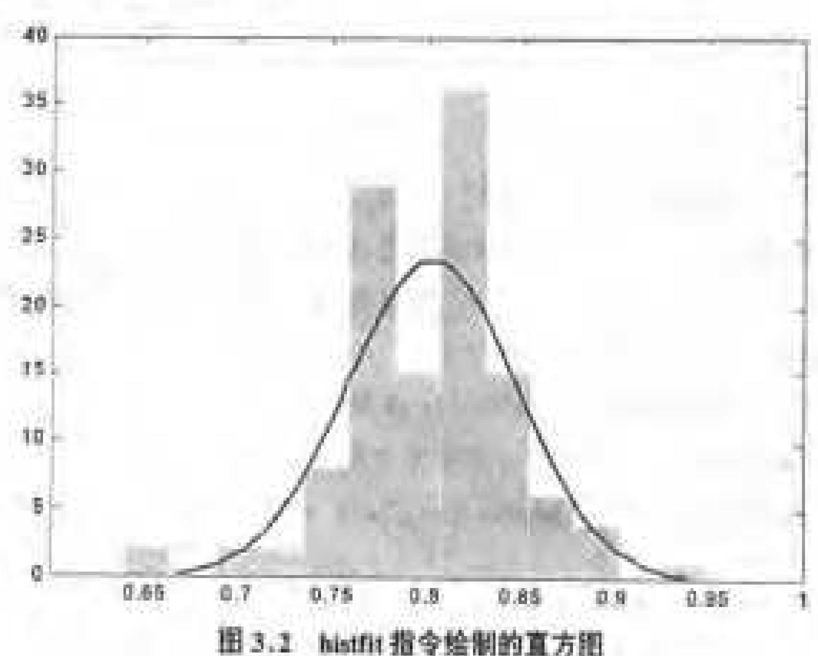


图3.2 histfit指令绘制的直方图

有时，人们常以频率分布表中的组中值为横坐标、以累积频率为纵坐标绘制累积频率折线(请读者用 plot 指令自行画出图形).

在应用中，可以根据频率直方图(累积频率折线图)了解变量的概率密度曲线(分布曲线)的大致形态，进而估计变量的分布类型. 在得出初步的结论后应继续通过分布参数的估计和分布拟合检验得出更为精细的结论.

对于离散型随机变量，一般在大样本条件下求样本数据的频率，画出不同数据点频率值的火柴杆图(或散点图)，通过对已知的离散分布的分布律图形作出变量分布形态的估计，供进一步分析参考，这里不再赘述.

下面举例说明直方图的应用.

【例 3.2】 用模拟试验的方法直观地验证定理 2.6 的结论.

假定变量 $X\sim N(60,25)$，用随机数生成的方法模拟对 X 的 500 次简单随机抽样，每个样本的容量为 16. 利用这 500×16 个样本数据直观地验证样本均值 $\overline{X}$ 的抽样分布为均值等于 60、方差等于 25/16 的正态分布，即 $\overline{X}\sim N(60,25/16)$.

① 用随机数生成的方法模拟简单随机抽样.

```
clear
x=[]; %生成一个存放样本数据的空表(维数可变的动态矩阵)
for byk=1:500 %循环控制，循环执行下面的指令 500 次，本例中相当于 500 次抽样
xx=normrnd(60,5,16,1); %生成一个来自 N(60, 25)的容量为 16 的样本(列向量)
x=[x,xx]; %将样本数据逐列存入数表 x，可从 Matlab 的变量浏览器(Workspace)中观察这个数表
end %循环重复标志
```

② 计算每一个样本的样本均值，得到 $\bar{x}_1$，$\bar{x}_2$，…，$\bar{x}_{500}$.

```
xmean=mean(x); %可从 Matlab 的变量浏览器中观察这 500 个数据
```

③ 绘制 500 个样本均值 $\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_{500}$数据的直方图. 如果直方图是单峰对称的，则可认定样本均值 $\bar{x}$ 的抽样分布是正态分布.

```
k=ceil(1.87*(length(x)-1)^(2/5)); %确定分组数
h=histfit(xmean,k); %绘制附正态参考曲线的数据 x̄1,x̄2,…,x̄500的直方图
set(h(1),'FaceColor','c','EdgeColor','w') %修饰，设置直方图线条颜色与填充色
```

上述指令的运行结果见图 3.3.

④ 用这 500 个样本均值数据验证 $\bar{x}$ 的均值等于 60，方差等于 $25/16=1.5625$.

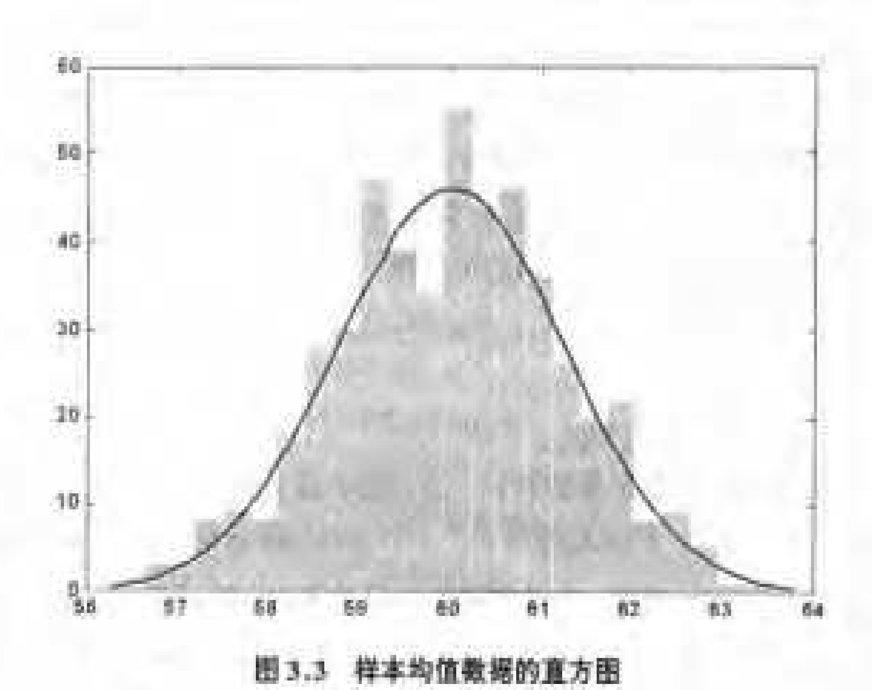


图 3.3 样本均值数据的直方图

```
M = mean(xmean) % 求 x̄1, x̄2, …, x̄500的均值，以此作为 Ex̄ 的近似值
V = var(xmean) % 求 x̄1, x̄2, …, x̄500的方差，以此作为 varx̄ 的近似值
```

上述指令的运行结果是：

```
M =
   60.0133
V =
   1.5649
```

上述结果表明，样本均值 $\bar{x}$ 的抽样分布是正态的，且用 $\bar{x}_1$，$\bar{x}_2$，…，$\bar{x}_{500}$的样本均值与样本方差近似 $\bar{x}$ 的数学期望与方差的效果较好，这就直观地验证了定理 2.6 的结论(更严谨的，应当进行分布拟合检验与参数检验，相关内容在第 4 章介绍)．

3.1.2 经验分布函数

定义 3.1 (经验分布函数) 设 X_1，X_2，…，X_n i.i.d. $\sim F(x)$，$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$是顺序统计量 $X_{(1)}$，$X_{(2)}$，…，$X_{(n)}$的观测值．令

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}\nu_i, & x_{(k)}\leqslant x<x_{(k+1)},\quad (k=1,2,\cdots,n-1)\\ 1, & x\geqslant x_{(n)},\end{cases}$$

其中 ν_i 为样本数据 $x\in[x_{(i)}, x_{(i+1)})$的频数，则称 $F_n(x)$为该样本的经验分布函数．

经验分布函数在 x 点的函数值其实就是样本观测值 $x\leqslant x_{(k)}$的累积频率．

经验分布函数是一个右连续的非降函数，且 $0 \leqslant F_n(x) \leqslant 1$.

经验分布函数具有分布函数的性质. 我们可以将经验分布函数理解为是以等概率取 $x_1, x_2, \cdots, x_n$ 的离散型随机变量的分布函数，其图像是一个非降右连续的阶梯函数. 经验分布函数在应用中十分重要，它可以用来描述总体分布函数的大致形状. 下面的 Гливенко 定理从理论上证明了这种应用的可靠性.

定理 3.1（Гливенко 定理） 设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim F(x)$，$F_n(x)$为样本的经验分布函数，则

$$P\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\} = 1.$$

证明涉及较多的概率极限定理的知识，感兴趣的读者请参见文献[7]. 在定理 3.1 中，$D_n = \sup\limits_{-\infty<x<+\infty} |F_n(x) - F(x)|$是对 $F_n(x)$和 $F(x)$在 x 的一切取值上的最大差异的衡量. $P\{\lim\limits_{n\to\infty} D_n = 0\} = 1$ 说明，当样本容量 n 足够大时，对一切 x，$F_n(x)$可以按给定的精确度接近 $F(x)$，这一事件发生的概率为 1. 因此，Гливенко 定理是数理统计用样本数据对变量的分布形态及分布参数进行推断的理论依据.

下面举例说明经验分布函数图像的 MATLAB 绘制及应用.

【例 3.3】 例 3.1 中 25MnSi 钢含硅量数据的经验分布函数.

经验分布函数是一种在大样本条件下估计变量分布形态的重要工具. 经验分布函数的图像与累积频率折线图在性质上是一致的，它们的主要区别在数据的分组上，经验分布函数处理得更为细腻.

应用中可以将经验分布函数图像与可能的分布类型的分布函数图像进行对比，得出关于变量分布形态的结论.

经验分布函数图像 MATLAB 绘图指令是 cdfplot，其输入参数为样本数据向量，有两个可选输出参数：第一个是图形句柄；第二个是关于样本数据的几个重要的统计量，包括样本最小值、最大值、均值、中值和标准差.

```
clear
load hgl
[h,stats] = cdfplot(hgl)
```

上述指令的运行结果如下：

```
h =
    154.0016
stats =
          min:  0.6400
          max:  0.9500
         mean:  0.8026
```

```
median:  0.8100
   std:  0.0450
```

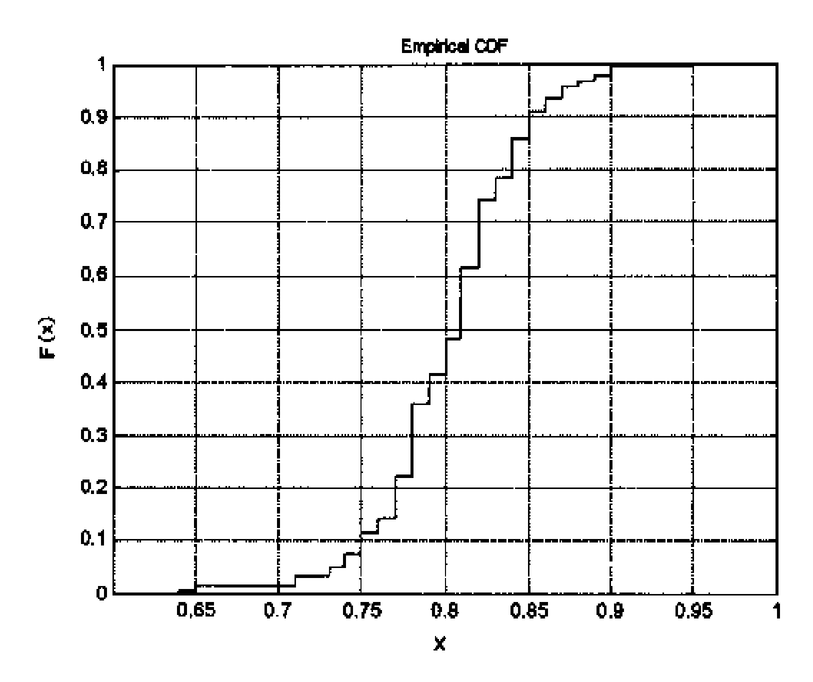


图3.4 120炉25MnSi钢含硅量数据的经验分布函数图像

由图3.4可以看出，样本经验分布函数图像上升速度较快，均值与中值接近，图像的S形状均衡对称，均值处函数值约为0.5. 这些特征表明，25MnSi钢的含硅量可能服从均值为0.8026、标准差为0.045的正态分布. 接下来，可以通过正态拟合检验进一步证实这种推测.

3.1.3 五数概括与box图

度量数据分布特征常用的统计量包括样本峰度、样本偏度和百分比分位数. 在2.2.2节已经给出它们的数学定义及相应的MATLAB函数. 下面对这几个概念作进一步的说明.

样本峰度 $KU=\frac{\nu_4}{\nu_2^2}$ 是对单峰分布曲线"峰的平坦程度"或者说"曲线在其峰值附近的陡峭程度"的度量. 对于样本峰度的定义，不同文献有所不同，一般定义为 $KU=\frac{\nu_4}{\nu_2^2}-3$，此时正态分布具有零峰度. 这里采用了MATLAB系统中样本峰度的定义，正态分布的峰度为3. 当变量的样本峰度大于3时，其密度曲线比正态分布密度曲线陡峭；当变量的样本峰度小于3时，其密度曲线比正态分布密度曲线平坦. 这里，$\nu_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-$

$\bar{x})^k(k>0)$ 是样本的 k 阶中心矩 B_k 的观测值.

样本偏度 $SK=\frac{\nu_3}{\nu_2^{3/2}}$ 是对变量的分布围绕其均值的对称情况的度量. 如果样本偏度等于 0, 则变量分布的形状是对称的(如正态分布); 如果样本偏度大于 0, 则变量分布的形状是右尾长, 变量取值的密度左边偏大, 称为正(或右)偏的; 如果样本偏度小于 0, 则变量分布的形状是左尾长, 变量取值的密度右边偏大, 称为负(或左)偏的.

样本的百分比分位数亦称为样本 p 分位数, 表示如下:

$$m_p=\begin{cases} x_{([np+1])}, & np\notin \mathbf{N}, \\ 0.5(x_{(np)}+x_{(np+1)}), & np\in \mathbf{N}, \end{cases}$$

其中 **N** 为正整数集. 关于样本的百分比分位数, 应用最多的是样本的 4 分位数 $Q_1=m_{0.25}$, $Q_2=m_{0.5}$ 和 $Q_3=m_{0.75}$, 分别称为第一 4 分位数、第二 4 分位数与第三 4 分位数, 它反映了有 1/4 的数据小于 Q_1, 有 1/4 的数据大于 Q_3, 而有一半数据介于 Q_1 与 Q_3 之间.

接下来给出这几个概念在估计变量分布形态方面的一种综合应用——**五数概括与 box 图**.

在统计应用中, 常用样本数据的最小值、是大值和 4 分位数来反映变量分布的信息, 称为五数概括, 而 box 图(箱线图)则是五数概括的图形化, 见图 3.5.

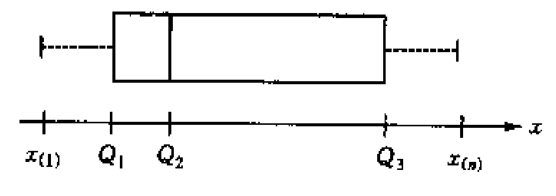


图 3.5 box 图示意

从 box 图可以看出样本数据的如下特征, 并可以此来推测变量的分布特点.

① 中心位置. 中位数 $Q_2=m_{0.5}$ 所在的位置即为样本数据的中心, 在 $[x_{(1)}, Q_2]$ 和 $[Q_2, x_{(n)}]$ 中各包含样本数据的一半.

② 散布情况. 全部样本数据位于 $[x_{(1)}, x_{(n)}]$ 内, 若将样本数据等分成四份的话, 那么在区间 $[x_{(1)}, Q_1]$, $[Q_1, Q_2]$, $[Q_2, Q_3]$ 和 $[Q_3, x_{(n)}]$ 内各占 1/4. 各区间较短时, 特别是 $[x_{(1)}, x_{(n)}]$ 与 $[Q_1, Q_3]$ 较短时, 表示样本较集中, 反之就较为分散.

③ 偏度. 如果矩形位于中间位置, 中位数又位于矩形的中间位置, 则分布较为对称, 否则是偏态分布. 如果矩形偏于左端(或右端), 中位数偏于矩形左端(或右端), 可知分布是正偏(或负偏), 此时右(左)尾较长. box 图偏度解读见图 3.6.

④ 离群值. 当矩形两端线段长度相差过大时, 表明长的一侧有特大(或特小)值, 称

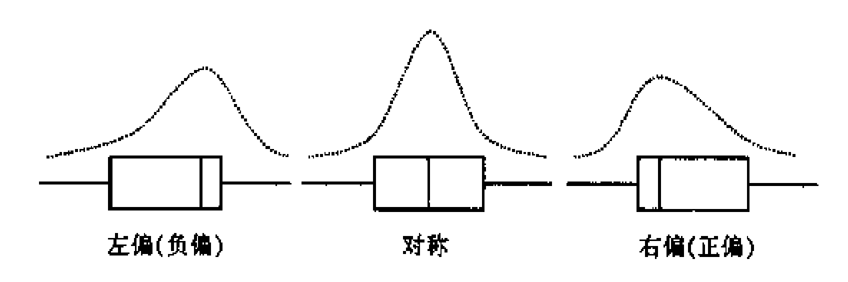


图3.6　box图偏度解读

为离群值，用“+”标记，而线段终于 $x_{(n-1)}$(或 $x_{(2)}$)，甚至终于 $x_{(n-2)}$(或 $x_{(3)}$)．

【例3.4】　设有两个教学班，各有30名同学，在数学课程上，A班用新教学方法组织教学，B班用传统方法组织教学，现得期末考试成绩如下．

A：82，92，77，62，70，36，80，100，74，64，63，56，72，78，68，65，72，70，58，92，79，92，65，56，85，73，61，71，42，89

B：57，67，64，54，77，65，71，58，59，69，67，84，63，95，81，46，49，60，64，66，74，55，58，63，65，68，76，72，48，72

试在同一坐标轴上画出相应的box图，并对两个班的成绩进行初步的分析比较．

MATLAB绘制box图的指令是boxplot．

```
clear
x=[82, 92, 77, 62, 70, 36, 80, 100, 74, 64, 63, 56, 72, 78, 68, 65, 72, 70,
58, 92, 79, 92, 65, 56, 85, 73, 61, 71, 42, 89; 57, 67, 64, 54, 77, 65, 71, 58, 59,
69, 67, 84, 63, 95, 81, 46, 49, 60, 64, 66, 74, 55, 58, 63, 65, 68, 76, 72, 48, 72];
boxplot(x') % boxplot 指令将输入矩阵的每一列视为一个变量(的样本数据)
```

上述指令的运行结果见图3.7．

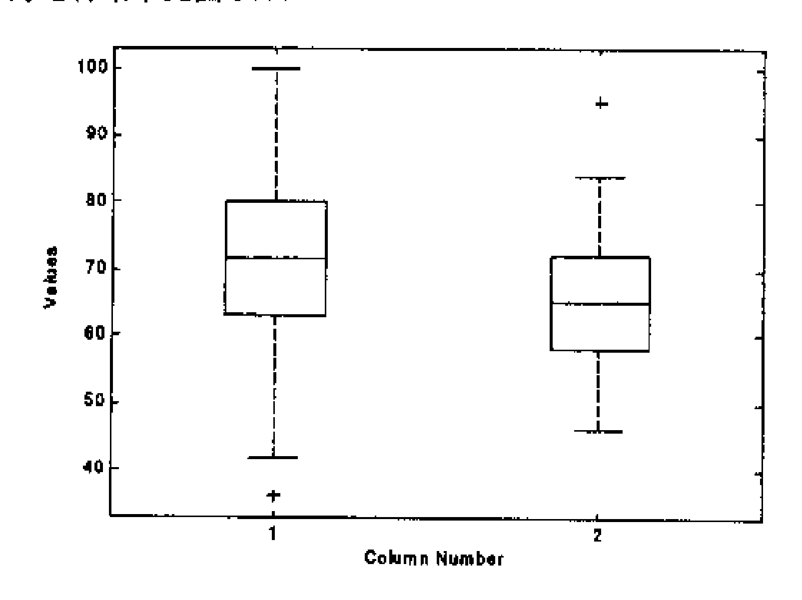


图3.7　两个班的成绩的box图

从图 3.7 中可以直观地看出，两个班的数学成绩的分布是正态(对称)的，A 班成绩较为分散(方差大)，B 班成绩则较集中(方差小). A 班成绩明显高于 B 班(均值比较，并且 A 班 25% 低分段上限接近 B 班中值线，A 班中值线接近 B 班 25% 高分段下限)，A 班的平均成绩约为 70 分(中值)，B 班约为 65 分(中值)，A 班有一名同学的成绩过低(离群)，而 B 班成绩优秀的只有一人(离群). 需要注意的是，从图 3.7 中我们不能得出新教学方法一定优于传统教学方法的结论，因为我们并不知道两个班级原有的数学基础是怎样的.

3.2 变量分布参数的估计

3.2.1 参数估计的方法

本小节讨论参数估计问题，即利用抽样信息来估计变量的分布参数或者参数的某个函数. 在参数估计问题中，我们总是假定变量具有已知的分布形式，未知的仅仅是一个或几个参数. 然而，变量的真分布完全由这些参数所决定，因此通过估计参数可以估计变量的真分布(有时我们仅仅需要估计这些参数).

设变量 X 的分布函数 $F(x;\theta)$ 的形式已知，θ 为待估参数(θ 是有限维向量). 为估计 θ，抽取 $X_1, X_2, \cdots, X_n$，构造出适当的统计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$，$\hat{\theta}$ 与 θ 有相同的维数和取值范围. 每当有了样本 $X_1, X_2, \cdots, X_n$ 的观测值，就代入函数 $\hat{\theta}(X_1, \cdots, X_n)$ 算出一个值，用来作为 θ 的估计值.

为着这样特定目的而构造的统计量 $\hat{\theta}$，叫做参数 θ 的估计量.

参数估计常用的方法是矩方法和极大似然法.

3.2.1.1 矩估计法

矩估计法是 K.Pearson 在 19 世纪提出来的，是一种基于简单的“替换”思想建立起来的估计方法.

在 Гливенко 定理的基础上可以证明，**样本矩 A_k 依概率收敛于变量 X 的 k 阶矩 $E(X^k)$，样本中心矩 B_k 依概率收敛于变量 X 的 k 阶中心矩 $E(X-E(X))^k$**. 因此，当样本容量 n 很大时，样本矩的观察值比较靠近变量的相应矩，就可以用样本矩去估计变量的相应矩. 这是矩估计法的基本数学原理.

例如，用样本 k 阶矩 $A_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$ 作为变量的 k 阶矩 $E(X^k)$ 的估计量，用样本 k 阶中心矩 $B_k = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$ 作为变量的 k 阶中心矩 $E((X-E(X))^k)$ 的估计量.

矩估计的一般方法是：设变量的分布函数含有 k 个未知参数 $\theta_1, \cdots, \theta_k$，那么它的前 k 阶矩 $\mu_1, \cdots, \mu_k$(必要时也可以是中心矩)一般都是这 k 个参数的函数，记为

$$\mu_i = g_i(\theta_1, \cdots, \theta_k) \quad (i=1,2,\cdots,k),$$

假如能从这 k 个方程中解出

$$\theta_j = h_j(\mu_1, \cdots, \mu_k) \quad (j=1,2,\cdots,k),$$

那么用诸 μ_i 的矩估计量 A_i 分别代替上式中的诸 μ_i，即可得诸 θ_j 的矩估计量

$$\hat{\theta}_j = h_j(A_1, \cdots, A_k) \quad (j=1,2,\cdots,k).$$

【例3.5】 设变量 X 的均值 μ 和方差 σ^2 都存在，μ 和 σ^2 均未知. 又设 $X_1, X_2, \cdots, X_n$ i.i.d.$\sim X$，求 μ 和 σ^2 的矩估计量.

解 因为变量 X 的分布中只含两个未知参数 μ 和 σ^2，故需求出变量 X 的一阶、二阶矩

$$\begin{cases} \mu_1 = E(X) = \mu, \\ \mu_2 = E(X^2) = \mu^2 + \sigma^2. \end{cases}$$

由矩估计法，用样本矩去替换总体矩，即令

$$\begin{cases} \mu = A_1, \\ \sigma^2 + \mu^2 = A_2, \end{cases}$$

解上述方程组，得 μ 和 σ^2 的矩估计量分别为

$$\begin{cases} \hat{\mu} = A_1 = \bar{X}, \\ \hat{\sigma}^2 = A_2 - A_1^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{cases}$$

本题也可以用样本的二阶中心阶 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 直接去估计总体的二阶中心阶 $D(X) = \sigma^2$.

矩估计法简便易行，使用时并不需要事先知道变量的分布. 但是，在变量分布类型已知的场合，矩估计法没有充分利用变量的分布所提供的信息. 一般场合下，矩估计量不具有唯一性. 如泊松分布参数的矩估计量既可以是样本均值，又可以是样本方差.

3.2.1.2 极大似然估计法

首先举例说明极大似然估计法的数学原理.

【例3.6】 设有甲、乙两个布袋，甲袋中有99个白球和1个黑球，乙袋中有1个白球和99个黑球. 由于某种原因已不能识别哪一个是甲袋，哪一个是乙袋. 你能否用统计的方法识别出来？

下面对这一问题进行数学描述与分析.

不妨设变量 X 表示袋中的白球数，则 $X \sim \begin{pmatrix} 1 & 99 \\ p & 1-p \end{pmatrix}$，$p$ 是未知的分布参数，其取值依赖于变量 X 代表的是甲袋中的白球数还是乙袋中的白球数. 显然，变量 X 代表的是甲袋中的白球数与 $p=99/100$ 是等价的，变量 X 代表的是乙袋中的白球数与 $p=1/100$ 是等价的.

我们可以通过抽样(任取一袋，从该袋中任取一球，观察其颜色)的方法来确定 $p=99/100$ 还是 $p=1/100$.

设事件 A 表示“取出的一袋为甲袋”，事件 B 表示“从袋子中取出的是白球”，则

$$P(A)=0.5,\quad P(B\mid A)=99/100,\quad P(B\mid \bar{A})=1/100.$$

假定取出的是白球. 在已知取出的是白球的条件下，判断该球来自甲袋还是乙袋的问题，可由贝叶斯公式，通过比较概率 $P(A\mid B)$和 $P(\bar{A}\mid B)$的大小来作出判断. 由于在一次试验中大概率事件容易发生，因此，若 $P(A\mid B)>P(\bar{A}\mid B)$，则该球来自甲袋；若 $P(A\mid B)<P(\bar{A}\mid B)$，则该球来自乙袋.

因为

$$P(A\mid B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B\mid A)}{P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})},$$

$$P(\bar{A}\mid B)=\frac{P(\bar{A}B)}{P(B)}=\frac{P(\bar{A})P(B\mid \bar{A})}{P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})},$$

这两个式子的分母相同，分子中 $P(A)=P(\bar{A})$，故其大小取决于 $P(B\mid A)$和 $P(B\mid \bar{A})$的大小，而 $P(B\mid A)$和 $P(B\mid \bar{A})$的取值恰好等于变量 X 的分布参数 p 的两个可能的取值. 这说明参数的取值同逆概率 $P(B\mid A)$与 $P(B\mid \bar{A})$之间的大小是相互决定的，即 $p=99/100$ 等价于 $P(A\mid B)>P(\bar{A}\mid B)$；$p=1/100$ 等价于 $P(A\mid B)>P(\bar{A}\mid B)$.

通过计算可知，$P(A\mid B)>P(\bar{A}\mid B)$，因此 $p=99/100$，即现在取出的这一袋是甲袋.

概括这里的思想方法，就可以得到极大似然估计法的数学原理——大概率原理：**大概率事件在一次试验中容易发生. 或者说，在一次试验中已经发生的事件具有较大的概率，而变量的分布参数有助于关于该变量的大概率事件的发生.**

接下来讨论参数的极大似然估计的方法.

设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$，并记变量 X 的概率分布律或概率密度函数为 $p(x;\theta_1,\theta_2,\cdots,\theta_k)$，其中 $\theta_1,\theta_2,\cdots,\theta_k$ 是变量X 的 k 个未知参数.

又设对样本$(X_1, X_2,\cdots,X_n)$进行一次观测得到样本值$(x_1,x_2,\cdots,x_n)$，这相当于 n 个相互独立的事件$\{X_1=x_1\}$，$\{X_2=x_2\}$，…，$\{X_n=x_n\}$在一次试验中同时发生，即事件$\{X_1=x_1, X_2=x_2, \cdots, X_n=x_n\}$应该有较大的概率值.

(1) X 是离散变量的情形

根据前述极大似然估计法的数学原理，可令

$$P\{X_1 = x_1,\ X_2 = x_2,\ \cdots,\ X_n = x_n\} = \prod_{i=1}^{n} P\{X_i = x_i\} = \prod_{i=1}^{n} p(x_i;\ \theta_1,\theta_2,\cdots,\theta_k)$$

达到最大值，此时对应的参数值 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$ 即为参数真值 $\theta_1,\theta_2,\cdots,\theta_k$ 的估计值.

(2) X 是连续变量的情形

对连续变量考虑概率 $P\{X_1=x_1,\ X_2=x_2,\ \cdots,\ X_n=x_n\}$ 是没有意义的. 因此，我们考虑随机点 $(X_1,\ X_2,\cdots,X_n)$ 落入以点 $(x_1,x_2,\cdots,x_n)$ 为顶点，$\Delta x_1,\Delta x_2,\cdots,\Delta x_n$ 为边长的 n 维矩形区域 G 内的概率，这个概率近似等于

$$P\{(X_1,\ X_2,\cdots,X_n)\in G\} = \prod_{i=1}^{n} p(x_i;\ \theta_1,\ \theta_2,\ \cdots,\ \theta_k)\cdot\prod_{i=1}^{n}\Delta x_i.$$

同理，可令这个概率达到最大值，此时对应的参数值 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$ 即为参数真值 $\theta_1,\theta_2,\cdots,\theta_k$ 的估计值.

注意到 $\Delta x_i(i=1,2,\cdots,n)$ 与 $\theta_1,\theta_2,\cdots,\theta_k$ 无关，使 $\prod_{i=1}^{n} p(x_i;\ \theta_1,\theta_2,\cdots,\theta_k)\prod_{i=1}^{n}\Delta x_i$ 达到是大值的点 $(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k)$ 与使 $\prod_{i=1}^{n} p(x_i;\ \theta_1,\theta_2,\cdots,\theta_k)$ 达到最大值的点相同，而后者在表达形式上连续型变量与离散型变量是一致的，因此给出下面的定义.

定义 3.2　称样本 $x_1,\ x_2,\ \cdots,\ x_n$ 的联合概率函数(概率分布律或概率密度函数)

$$L(\theta) = L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta) = \prod_{i=1}^{n} p(x_i;\ \theta)$$

为参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_k)$ 的似然函数.

设 Θ 为参数 θ 所有可能的取值范围，称为参数空间. 若存在统计量 $\hat{\theta}\in\Theta$，使得

$$L(x_1,\cdots,x_n;\hat{\theta}) = \max_{\theta\in\Theta} L(x_1,\cdots,x_n;\theta),$$

则称 $\hat{\theta}$ 是参数 θ 的极大似然估计量(Maximum Likelihood Estimate, MLE).

求似然函数 $L(\theta)$ 的极大值一般情况下要先求其驻点，涉及导数运算. 由于似然函数 $L(\theta)$ 的数学表达式往往是积与幂的结构，其导数运算会比较冗繁，不方便求驻点，而对数函数 $\ln x$ 是 x 的单调增函数，因此对数似然函数 $\ln L(\theta)$ 与似然函数 $L(\theta)$ 在同一点处取得最大值. 又对数能够将积运算转化为和运算，将幂运算转化为积运算，从而使似然函数 $L(\theta)$ 的数学表达式线性化，方便导数与求驻点运算. 于是，通常情况下应当先将似然函数 $L(\theta)$ 转化为对数似然函数 $\ln L(\theta)$，然后再求驻点.

【例 3.7】　求事件 A 发生的概率 p 的极大似然估计.

解 令 $X=\begin{cases}1, & \omega\in A,\\ 0, & \omega\notin A,\end{cases}$ 其中 $\omega\in A$ 表示事件 A 发生，则 X 的概率函数为

$$p(x;p)=p^x(1-p)^{1-x}\quad(x=0,1),$$

故参数 p 的似然函数为

$$L(p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)},$$

对数似然函数为

$$\ln L(p)=\Big(\sum_{i=1}^{n}x_i\Big)\ln p+\Big(n-\sum_{i=1}^{n}x_i\Big)\ln(1-p).$$

对 p 求导数，令导数为 0，就有

$$\frac{\mathrm{d}\ln L(p)}{\mathrm{d}p}=\frac{1}{p}\Big(\sum_{i=1}^{n}x_i\Big)-\frac{1}{1-p}\Big(n-\sum_{i=1}^{n}x_i\Big)=0,$$

解得 $\ln L(p)$ 的驻点为

$$p=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

又在驻点处有

$$\frac{\partial^2\ln L(p)}{\partial p^2}=\frac{-n}{p(1-p)}<0,$$

所以，驻点即为极大值点，即 p 的极大似然估计为 $\hat{p}=\bar{x}$.

【例 3.8】 设 $X\sim N(\mu,\sigma^2)$，求 μ 和 σ^2 的极大似然估计.

解 正态总体 $N(\mu,\sigma^2)$ 的密度函数是 $\frac{1}{\sqrt{2\pi}\sigma}\cdot\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$，则似然函数为

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\cdot\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\cdot\mathrm{e}^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}.$$

将其取对数，并令关于 μ，σ^2 的一阶导数为零，则得

$$\frac{\partial\ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0,$$

$$\frac{\partial\ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0.$$

解此关于 μ，σ^2 的方程组，得驻点

$$\mu=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

又可求得对数似然函数的二阶导函数矩阵是非正定矩阵，因此驻点处即为似然函数的极

大值点处，并将 μ 的样本表达式代入 σ^2 的驻点表达式，得 μ 与 σ^2 的极大似然估计为

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

【例3.9】 设 $X\sim U(a,b)$，求 a，b 的极大似然估计.

解 由于 X 的密度函数为

$$f(x)=\begin{cases}(b-a)^{-1}, & a\leqslant x\leqslant b,\\ 0, & \text{其他},\end{cases}$$

故似然函数为

$$L(a,b)=\begin{cases}(b-a)^{-n}, & a\leqslant x_i\leqslant b\quad (i=1,2,\cdots,n),\\ 0, & \text{其他},\end{cases}$$

显然

$$\frac{\partial L(a,b)}{\partial a}=\frac{n}{(b-a)^{n+1}}>0,$$

即 $L(a,b)$ 是关于 a 的单调增函数，因此，为使 $L(a,b)$ 达到最大，应使 a 最大. 同理

$$\frac{\partial L(a,b)}{\partial b}=-\frac{n}{(b-a)^{n+1}}<0,$$

即 $L(a,b)$ 是关于 b 的单调减函数，因此，为使 $L(a,b)$ 达到最大，应使 b 最小.

又对于任意的样本观测值 $x_1,x_2,\cdots,x_n$，恒有 $a\leqslant x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}\leqslant b$，于是，$a$，$b$ 的极大似然估计分别为 $\hat{a}=x_{(1)}$，$\hat{b}=x_{(n)}$.

根据前面几个例题的讨论，可以概括出求极大似然估计值的一般步骤：

① 明确变量的分布律或密度函数；

② 写出似然函数 $L(\theta)$；

③ 求似然函数 $L(\theta)$ 的最大值点，得 $\hat{\theta}_{\mathrm{MLE}}$；

④ 应用问题中，将样本数据代入 $\hat{\theta}_{\mathrm{MLE}}$ 求出具体的估计值.

值得注意的是，求解对数似然方程组是在假定其可导并且导数变号的基础上的，如例3.7和例3.8. 若不满足这一条件，需针对似然函数 $L(\theta_1,\theta_2,\cdots,\theta_k)$ 的单调性，利用极大似然估计的基本原理直接进行 $L(\theta_1,\theta_2,\cdots,\theta_k)$ 的最大值问题的讨论，如例3.9.

极大似然估计量有一个简单而有用的性质：设 θ 的函数 $g=g(\theta)$ 是 Θ 上的实值函数，且有唯一反函数. 如果 $\hat{\theta}$ 是 θ 的极大似然估计量，则 $g(\hat{\theta})$ 也是 $g(\theta)$ 的极大似然估计量. 这个性质称为极大似然估计的不变性. 根据这一性质可以使一些复杂结构的参数的极大似然估计问题简单化.

极大似然估计法是在变量分布类型已知的情况下使用的一种参数估计方法. 一般地，用极大似然法所得的估计的性质比用矩估计法所得的要好，故通常多用极大似然

法.

MATLAB 进行极大似然估计的函数为 mle. 其调用格式灵活多样(详见附录 B), 这里仅介绍一种最基本的调用方法:

```
[phat, pci] = mle(data, 'distribution', dist, 'alpha', a, 'ntrials', n)
```

其中, 输出参数 phat 是指定分布的参数的极大似然估计值(多参数时为行向量), pci 是参数的区间估计的置信上限和下限(与参数对应的二维列向量, 可以缺省). 输入参数 data 是样本数据向量(不可缺省). 引用参数'distribution'及其取值 dist 设置变量的分布类型(应用中 dist 要用具体的分布名称字符串替换并用单引号引起), 二者要成对出现(可以同时缺省, 缺省时分布类型默认为正态分布). 引用参数'alpha'及其取值 a 设置区间估计的显著性水平, 二者要成对出现(可以同时缺省, 缺省时默认为 0.05, 即置信水平为 0.95). 引用参数'ntrials'及其取值 n 仅在分布类型为二项分布时引用(对于其他分布可以缺省), 设置二项分布中试验的次数.

【例 3.10】 通常, 引力常数的测定值服从均值为 μ、标准差为 σ 的正态分布. 某人在实验中使用金球测定引力常数, 6 次测定观察值为: 6.683, 6.681, 6.676, 6.678, 6.679, 6.672. 试用极大似然估计法对未知参数 μ 和 σ 作出估计.

解 用 mle 函数进行计算.

```
clear
x=[6.683  6.681  6.676  6.678  6.679  6.672];
phat = mle(x, 'distribution', 'norm', 'alpha', 0.05)
```

上述指令的运行结果是:

```
phat =
       6.6782     0.0035
```

即金球测定的 μ 估计值为 6.6782, σ 的估计值为 0.0035. 其实, 此例计算中 mle 函数的调用可以简化为 p= mle(x).

3.2.2 估计量的性能分析

在参数估计问题中, 在可选择的估计量中哪个更好, 如何评价和控制估计误差, 这是除参数估计的方法外必须回答的两个问题. 本小节讨论第一个问题, 在下一小节讨论另一个问题.

在分析和评价估计量性能的时候, 常用的准则包括无偏性准则、均方误差准则和相合性准则.

3.2.2.1 无偏性准则

估计量是随机变量, 对于不同的样本值会得到不同的估计值. 我们希望估计值在未

知参数真值附近摆动，而它的期望值等于未知参数的真值．这就导致无偏性这个标准．

定义 3.3（无偏估计） 设$\hat{\theta}(X_1,\cdots,X_n)$是变量 X 的未知的一维参数θ的估计量，若 $E(\hat{\theta})=\theta$，则称$\hat{\theta}$为 θ 的无偏估计．否则称为有偏估计．

定义 3.4（渐近无偏估计） 设$\hat{\theta}(X_1,\cdots,X_n)$是变量 X 的未知的一维参数θ的有偏估计量，但是 $\lim\limits_{n\to\infty}E(\hat{\theta})=\theta$，则称$\hat{\theta}$为 θ 的渐近无偏估计．

下面，不加证明地列举出关于无偏性的几个重要结论．

① 无论变量 X 服从何种分布，样本的 k 阶原点矩$A_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^k(i=1,2,\cdots,n)$是变量 X 的 k 阶原点矩$E(X^k)$的无偏估计．自然，$\bar{X}$ 是$E(X)$的无偏估计．

② 无论变量 X 服从何种分布，样本(修正)方差 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$是变量 X 的方差 σ^2 的无偏估计．

③ 样本方差(二阶中心矩)B_2 不是变量的方差 σ^2 的无偏估计，但是 $\lim\limits_{n\to\infty}E(B_2)=\sigma^2$，所以 B_2 是 σ^2 的渐近无偏估计．

④ 样本标准差 $S=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}$ 不是变量 X 的标准差σ的无偏估计．但是，在变量的正态性假设下，可将样本标准差修正为$\hat{\sigma}_S=C_nS$，$\hat{\sigma}_S$ 是σ 的无偏估计，其中 $C_n=\sqrt{\dfrac{n-1}{2}}\dfrac{\Gamma\left(\dfrac{n-1}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)}$称为正态标准差的无偏系数．由于 $\lim\limits_{n\to\infty}C_n=1$，所以 S 是σ 的渐近无偏估计．

无偏性准则是对估计量的一个朴素要求．无偏性估计的统计意义是指估计量不产生系统性的偏差．例如，用样本均值 $\bar{X}$ 作为变量均值μ 的估计时，由于 $\bar{X}$ 是随机变量，故在一次估计中 μ 的实现值与其真值之间存在偏差$\bar{X}-\mu$．这种偏差是随机的，虽无法说明一次估计所产生的偏差，但是对同一统计问题大量重复使用 $\bar{X}$ 估计μ 时，实际产生的偏差 $\bar{X}-\mu$ 随机地在 0 的周围波动，不会产生系统的 $\bar{X}$ 偏大(小)于 μ 的情况．

渐近无偏是指估计量存在系统性的偏差，但是这种系统性偏差随着样本容量的增加而趋向于消失．

3.2.2.2 均方误差准则

如果在样本容量 n 相同的情况下，$\hat{\theta}_1$ 的观察值较$\hat{\theta}_2$ 的观察值更密集在真值 θ 的附近，我们就认为用$\hat{\theta}_1$ 对 θ 进行的估计优于用$\hat{\theta}_2$ 对 θ 进行的估计．

定义 3.5 (均方误差准则) 设$\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$是变量 X 的未知的一维参数θ的估计量，称 $MSE\ \hat{\theta}=E(\hat{\theta}-\theta)^2$ 为估计量$\hat{\theta}$的均方误差. 对于参数 θ 的任意两个估计量$\hat{\theta}_1$和$\hat{\theta}_2$，若 $MSE\ \hat{\theta}_1\leqslant MSE\ \hat{\theta}_2$，且在参数空间中至少有一个 θ_0，使不等式中的小于号"<"严格成立，则称在均方误差意义下$\hat{\theta}_1$是优于$\hat{\theta}_2$的估计.

定理 3.2 (均方误差的分解定理) $MSE\ \hat{\theta}=\mathrm{Var}(\hat{\theta})+[E(\hat{\theta})-\theta]^2$.

事实上

$$
\begin{aligned}
MSE\ \hat{\theta}&=E((\hat{\theta}-\theta)^2)=E((\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2)\\
&=E((\hat{\theta}-E(\hat{\theta}))^2)+2E((\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta))+[E(\hat{\theta})-\theta]^2.
\end{aligned}
$$

由于

$$E((\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta))=0,$$

所以

$$MSE\ \hat{\theta}=\mathrm{Var}(\hat{\theta})+[E(\hat{\theta})-\theta]^2.$$

若$\hat{\theta}$是 θ 的无偏估计，则 $MSE\ \hat{\theta}=\mathrm{Var}(\hat{\theta})$.

一个参数往往有不止一个无偏估计. 由均方误差的分解定理不难理解，无偏估计以方差小者为好.

定义 3.6 (最小方差无偏估计) 设$\hat{\theta}^*(X_1,X_2,\cdots,X_n)$是变量 X 的未知参数θ的一个估计量，若$\hat{\theta}$满足：

① $E(\hat{\theta}^*)=\theta$，即$\hat{\theta}^*$为 θ 的无偏估计，

② $\mathrm{Var}(\hat{\theta}^*)\leqslant\mathrm{Var}(\hat{\theta})$，$\hat{\theta}(X_1,X_2,\cdots,X_n)$是 θ 的任意一个无偏估计，

则称$\hat{\theta}^*$为 θ 的最小方差无偏估计(也称最佳无偏估计).

请注意下面几个关于最小方差无偏估计的结论：

① 最小方差无偏估计可能存在，也可能不存在；

② 对于正态变量 X，样本均值 $\overline{X}$ 和样本方差S^2是 μ 和σ^2 的最小方差无偏估计；

③ 极大似然估计往往是均方误差最小的估计.

均方误差准则是最为常用的估计量性能评价准则，可以这样理解它的统计意义：设$\hat{\theta}$为 θ 的一个估计，由于估计量是随机变量，故在一次估计中 θ 的实现值与其真值之间存在偏差$\hat{\theta}-\theta$. 我们希望这种偏差尽可能的小. 但是由于偏差是随机变量，因此，不能根据一次估计时偏差$\hat{\theta}-\theta$的大小来判断估计的优劣，而应根据对同一个参数 θ 用同一个估计量$\hat{\theta}$进行的多次估计的"平均偏差"来判断. 为避免求平均偏差时$\hat{\theta}-\theta$的正负值相

互抵消，我们使用$(\hat{\theta}-\theta)^2$表示一次估计中的(平方)误差. 于是，$MSE\,\hat{\theta}_1 \leqslant MSE\,\hat{\theta}_2$表明多次用估计$\hat{\theta}_1$和$\hat{\theta}_2$去估计$\theta$时，$\hat{\theta}_1$的观察值较$\hat{\theta}_2$的观察值更密集在真值$\theta$的附近. 换句话说，均方误差准则说明，当使用不同的估计量$\hat{\theta}_1$和$\hat{\theta}_2$去估计θ时，其均方误差越小，估计的效果越好；反之，均方误差越大，估计的效果越差.

3.2.2.3 相合性准则

无偏性准则和均方误差准则是在样本容量n固定的情形下讨论估计量优劣的. 设变量$X \sim F(x)$，$\hat{F}_n(x)$为样本的经验分布函数，由 Гливенко 定理

$$P\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|\hat{F}_n(x)-F(x)|=0\}=1,$$

当样本容量n趋向于无穷时，样本的经验分布函数以概率1一致收敛于变量的分布函数. 也就是说，当样本容量n趋向于无穷时，样本中包含的关于变量分布的信息不断增加，以致充分到可以将变量分布刻画到任意精确的程度. 因此，我们有理由要求，一个“好的”估计量，当样本容量n趋向于无穷时，在一定的数学意义下收敛于被估参数.

定义 3.7 (相合估计) 设$\hat{\theta}(X_1,X_2,\cdots,X_n)$为参数$\theta$的估计量，若对任意的$\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|\hat{\theta}-\theta|\geqslant\varepsilon\}=0,$$

而且这对θ的一切可能取的值都成立，则称$\hat{\theta}$是参数θ的一个相合估计.

相合性准则是对一个估计量最基本的要求. 它说明，随着样本容量的增大，一个“好的”估计量$\hat{\theta}$应该越来越靠近参数θ的真值，使绝对偏差$|\hat{\theta}-\theta|$较大的概率越来越小. 如果一个估计量没有相合性，那么，不论样本取多大，我们也不可能把未知参数估计到预定的精度. 这种估计量显然是不可取的.

下面，不加证明地列举出关于相合估计的几个重要结论.

① 相合估计具有不变性. 即当$\hat{\theta}_1$，$\hat{\theta}_2$，…，$\hat{\theta}_k$分别是θ_1，θ_2，…，θ_k的相合估计时，若$g(\theta_1,\theta_2,\cdots,\theta_k)$为连续函数，则$g(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k)$是$g(\theta_1,\theta_2,\cdots,\theta_k)$的相合估计.

② 样本的k阶原点矩$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$是变量$X$的$k$阶原点矩$E(X^k)$的相合估计，故样本均值$\bar{X}$是变量均值$\mu$的相合估计.

③ 样本的二阶中心矩$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$是变量$X$的方差$\sigma^2$的相合估计.

④ 样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$是变量的方差$\sigma^2$的相合估计，样本标准差

$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 是变量的标准差 σ 的相合估计.

⑤ 事件发生的频率是其概率的相合估计.

⑥ 极大似然估计量往往具有相合性.

3.2.3 估计误差的评价与控制

在参数估计的应用中，一个重要的问题是对估计误差的评价与控制. 为此，先引进抽样误差的概念.

定义 3.8 (抽样误差) 设 θ 为总体 X 的未知参数，$\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$是 θ 的无偏估计量，则称 $\varepsilon=|\theta-\hat{\theta}|$ 为估计量的**绝对抽样误差**，称 $\rho=\frac{\theta}{\hat{\theta}}$ 为估计量的**相对抽样误差**.

一个无偏估计量的抽样误差不是由于估计量自身的构造产生的，而是由于抽样方式和样本容量的原因形成的，是参数估计中不是错误的"错误"，在参数估计中产生抽样误差是不可避免的. 因此对参数估计的抽样误差进行分析和控制是参数估计理论与实践所必须的.

下面举例来说明估计误差的评价与控制原理.

【例 3.11】 讨论用样本均值 $\overline{X}$ 估计变量均值 μ 时估计误差的评价与控制.

由 3.2.2 节的讨论知道，在对变量均值 μ 进行统计估计时，样本均值 $\overline{X}$ 是 μ 的相合的、最小方差的无偏估计量. 即便如此，由于随机性的影响，必然会产生一定的估计误差，用 $|\overline{X}-\mu|$ 表示所产生的误差(绝对误差). 自然，我们希望能够将误差控制在一个可以接受的范围内，用 $\varepsilon(>0)$表示这种限度，称为**绝对误差限或边际误差**，即要求 $|\overline{X}-\mu|\leqslant\varepsilon$.

也是由于随机性的影响，对于指定的 ε，我们无法保证对任何一次抽样都有 $|\overline{X}-\mu|\leqslant\varepsilon$.

于是，随机性的问题还需要随机性的方法来回答，我们不去预先指定边际误差 ε，转而考虑采用"**大概率保证下的误差控制策略**"，即预先约定用 $\overline{X}$ 估计 μ 的可靠性概率 β，称之为用 $\overline{X}$ 估计 μ 的**置信水平**，习惯上记 $\beta=1-\alpha$，并称 α 为用 $\overline{X}$ 估计 μ 的**风险概率或显著性水平**，那么此时产生的边际误差 ε 是多少? 用数学语言描述就是：指定 $\beta\in(0,1)$，求使 $P\{|\overline{X}-\mu|\leqslant\varepsilon\}\geqslant1-\alpha$ 成立的 ε.

称 $P\{|\overline{X}-\mu|\leqslant\varepsilon\}\geqslant1-\alpha$ 为用估计量 $\overline{X}$ 估计 μ 时的**误差评价与控制准则**.

这属于概率问题的反问题，解决这类问题的基本思路是求已知分布的分位数. 将准则式变形为 $P\{-\varepsilon\leqslant\overline{X}-\mu\leqslant\varepsilon\}\geqslant1-\alpha$，于是，为确定 ε 的值，需先获得 $\overline{X}-\mu$ 的概率

分布的信息.

在关于统计量及其抽样分布问题的讨论中知道，$\bar{X}-\mu$ 的概率分布需要区分如下三种情形：

① 正态变量且 $\mathrm{Var}(X)=\sigma^2$ 已知，$U=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$（小样本应用）；

② 正态变量但 $\mathrm{Var}(X)=\sigma^2$ 未知，$T=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$（小样本应用）；

③ 任意变量，$U^*=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\overset{L}{\sim}N(0,1)$（大样本应用）.

由于 μ 未知，所以对 U，T 和 U^* 不称为统计量，而称为**枢轴量**. 下面在正态变量但 σ^2 未知的假定下对准则式进行**枢轴变换**，得 $P\left\{-\dfrac{\varepsilon}{S/\sqrt{n}}\leqslant T\leqslant\dfrac{\varepsilon}{S/\sqrt{n}}\right\}\geqslant 1-\alpha$. 通常将 α 等分为上、下双侧显著性水平各 $\alpha/2$，即

$$P\left\{T\leqslant-\frac{\varepsilon}{S/\sqrt{n}}\right\}\leqslant\frac{\alpha}{2},\quad P\left\{T\geqslant\frac{\varepsilon}{S/\sqrt{n}}\right\}\geqslant\frac{\alpha}{2},$$

于是

$$\frac{\varepsilon}{S/\sqrt{n}}=t_{1-\alpha/2}(n-1),\quad -\frac{\varepsilon}{S/\sqrt{n}}=t_{\alpha/2}(n-1)=-t_{1-\alpha/2}(n-1),$$

解得

$$\varepsilon=t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

即

$$P\left\{\bar{X}-t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\leqslant\mu\leqslant\bar{X}+t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right\}\geqslant 1-\alpha.$$

上式表明，用 $\bar{X}$ 估计 μ 时，**假如进行了 100 次重复估计，可以保证至少有 $100(1-\alpha)$ 次估计的误差不超过 $t_{1-\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}$，即 $|\bar{X}-\mu|\leqslant t_{1-\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}$，而作出这个判断犯错误的概率是 α.**

需要指出的是，由于枢轴量构造的原因，在有关方差的估计问题中需要进行的是相对误差的分析.

【例 3.12】 讨论用样本方差 S^2 估计变量方差 σ^2 时估计误差的评价与控制.

同样，由 3.2.2 节的讨论知道，S^2 是 σ^2 的相合的、无偏估计量，又已知在变量的正态性背景下，$\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，于是，用 S^2 估计 σ^2 时估计误差的评价与控制分析采用相对抽样误差的概念. 由于 $E(S^2)=\sigma^2$，所以 $1-\rho\leqslant\dfrac{\sigma^2}{S^2}\leqslant 1+\rho$，$\rho$ 是一个小的正数. 为讨论方便，记为

$$\rho_{-} \leqslant \frac{\sigma^2}{S^2} \leqslant \rho_{+} \quad (\rho_{-} < \rho_{+}),$$

称 $\rho_{-}(\rho_{+})$ 为**相对误差下(上)限**.

同样采用"**大概率保证下的误差控制策略**", 不去预先指定相对误差限 ρ_{-} 和 ρ_{+}, 而是指定估计的置信水平 $1-\alpha$, 求使准则式

$$P\left\{\rho_{-} \leqslant \frac{\sigma^2}{S^2} \leqslant \rho_{+}\right\} \geqslant 1-\alpha$$

成立的 ρ_{-} 和 ρ_{+}. 由于 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 对准则式进行枢轴变换, 得

$$P\left\{\frac{n-1}{\rho_{+}} \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \frac{n-1}{\rho_{-}}\right\} \geqslant 1-\alpha,$$

于是

$$\frac{n-1}{\rho_{-}} = \chi^2_{1-\alpha/2}(n-1), \quad \frac{n-1}{\rho_{+}} = \chi^2_{\alpha/2}(n-1),$$

解得

$$\rho_{-} = \frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}, \quad \rho_{+} = \frac{n-1}{\chi^2_{\alpha/2}(n-1)},$$

即

$$P\left\{\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)} S^2 \leqslant \sigma^2 \leqslant \frac{n-1}{\chi^2_{\alpha/2}(n-1)} S^2\right\} \geqslant 1-\alpha.$$

上式表明, 用 S^2 估计 σ^2 时, **假如进行了 100 次重复估计, 可以保证至少有 $100(1-\alpha)$ 次估计的相对误差不低于** $\rho_{-} = \frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}$, **不高于** $\rho_{+} = \frac{n-1}{\chi^2_{\alpha/2}(n-1)}$, **即** $\rho_{-} S^2 \leqslant \sigma^2 \leqslant \rho_{+} S^2$, **而作出这个判断犯错误的概率是** α.

从例 3.11 的讨论中可以看到, 在对估计误差进行评价与控制分析的过程中, 我们得到了一个由统计量构造的随机区间

$$\left[\bar{X} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right],$$

用 $\bar{X}$ 估计 μ 的问题转化为用这个区间俘获 μ 的问题.

从例 3.12 的讨论中可以看到, 在对估计误差进行评价与控制分析的过程中, 我们也得到了一个由统计量构造的随机区间

$$\left[\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)} S^2,\ \frac{n-1}{\chi^2_{\alpha/2}(n-1)} S^2\right],$$

用 S^2 估计 σ^2 的问题转化为用这个区间俘获 σ^2 的问题.

上述关于参数估计的误差评价与控制分析在方法上具有一般性. 通常, 人们根据分析结果的表现形式而将参数估计的误差评价与控制分析过程称为参数的区间估计.

定义 3.9 (区间估计) 设变量 X 的概率分布为 $F(x;\theta)$，其中 θ 是未知的分布参数，参数空间为 Θ，$X_1,X_2,\cdots,X_n$ 是来自变量 X 的样本，$\hat{\theta}_L=\hat{\theta}_L(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_U=\hat{\theta}_U(X_1,X_2,\cdots,X_n)$ 是两个统计量，$\hat{\theta}_L<\hat{\theta}_U$. 对于给定的一个很小的正数 $\alpha(0<\alpha<1)$ 及任意的 $\theta\in\Theta$，若

① $P\{\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\}\geqslant1-\alpha$，则称随机区间 $[\hat{\theta}_L,\ \hat{\theta}_U]$ 为置信水平为 $1-\alpha$ 的(双侧)**置信区间**，$\hat{\theta}_L,\hat{\theta}_U$ 分别称为(双侧)**置信下限**和(双侧)**置信上限**；

② $P\{\hat{\theta}_L\leqslant\theta\}\geqslant1-\alpha$，则称随机区间 $[\hat{\theta}_L,\ +\infty)$ 为置信水平为 $1-\alpha$ 的**下侧置信区间**，$\hat{\theta}_L$ 称为**单侧置信下限**；

③ $P\{\theta\leqslant\hat{\theta}_U\}\geqslant1-\alpha$，则称随机区间 $(-\infty,\ \hat{\theta}_U]$ 为置信水平为 $1-\alpha$ 的**上侧置信区间**，$\hat{\theta}_U$ 称为**单侧置信上限**.

区间估计的实质是在事先对估计结果的可信程度作出承诺的情况下，给出参数估计值的同时对估计的抽样误差也作出了相应的判断.

在区间估计中，置信水平 $1-\alpha$ 刻画了所求得的随机区间 $[\hat{\theta}_L,\ \hat{\theta}_U]$ 俘获参数 θ 的可信程度，即区间 $[\hat{\theta}_L,\ \hat{\theta}_U]$ 有 $100(1-\alpha)\%$ 的机会俘获参数 θ. α 称为估计的**风险水平**(或**显著性水平**)，它刻画的是断定区间 $[\hat{\theta}_L,\ \hat{\theta}_U]$ 可俘获参数 θ 的误判概率. 置信区间的平均长度 $E(|\hat{\theta}_U-\hat{\theta}_L|)$ 表达了区间估计的精确度.

自然，我们希望得到在较大的置信水平下具有较高精确度的区间估计. 也就是说，要求估计的

① 置信水平 $1-\alpha$ 尽可能高，即概率 $P\{\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\}$ 要尽可能大；

② 精确度尽可能高，即区间的平均长度 $E(|\hat{\theta}_U-\hat{\theta}_L|)$ 尽可能小.

但是，理论分析表明这是一个两难问题：在固定样本容量的条件下，提高估计的精确度会使估计的置信水平下降，而提高估计的置信水平会使估计的精确度下降.

例如，估计一个人的体重在某一区间内，例如在[60,70](单位：kg)内，我们要求该估计量可靠，即有很大的把握此人的体重在这个范围内. 同时，也要求这个区间不能太长，区间长了，可靠度提高了，但精度也差了，这是一对矛盾.

在实际应用中，人们一般是在保证可靠的条件下尽量提高精度. 即以置信水平为主导，首先要保证估计结论的可信程度，然后再设法提高精确度. 换句话说，所谓"大概率保证下的误差控制策略"的意义是，预设估计的置信水平 $1-\alpha$，寻求适当的 $\hat{\theta}_L$，$\hat{\theta}_U$，使估计的精确度尽可能的高，即 $E(|\hat{\theta}_U-\hat{\theta}_L|)$ 尽可能的小. 在实际应用中，**若一定的置信水平下估计的精确度不满足要求，则唯一的解决办法就是增加样本容量**.

需要指出的是，置信区间不是唯一的．对同一个参数，我们可以构造许多置信区间（如对边际概率即显著性水平 α 进行不同的分配则可得不同的置信区间）．通常总是希望置信区间尽可能短．可以证明，在枢轴量的概率密度为单峰且对称的情形，如正态分布、t 分布，对于给定的样本容量和置信水平，对称于原点的置信区间的长度为最短．即使在概率密度不对称的情形，如 χ^2 分布、F 分布，习惯上仍取对称的分位数来计算未知参数的置信区间．

关于区间估计的方法，应用中要区分正态变量分布参数的小样本估计与非正态变量分布参数的大样本估计两种方法．下面忽略方法的推导过程，给出方法要点．

(1) 正态变量分布参数的小样本估计方法

正态变量分布参数的小样本估计方法见表 3.2．

表 3.2　　正态变量分布参数的小样本估计方法一览表

背景	参数	估计量	枢轴量及其分布	误差限	置信区间
$X\sim N(\mu,\sigma^2)$	μ	$\bar{X}$	$U=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$，$\sigma^2$ 已知	$\varepsilon=u_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	$\bar{X}\pm\varepsilon$
			$T=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$，$\sigma^2$ 未知	$\varepsilon=t_{1-\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}$	$\bar{X}\pm\varepsilon$
	σ^2	S^{*2} 或 S^2	$\chi^2=\dfrac{nS^{*2}}{\sigma^2}\sim\chi^2(n)$，$\mu$ 已知	$\rho_-=\dfrac{n}{\chi^2_{1-\alpha/2}(n)}$ $\rho_+=\dfrac{n}{\chi^2_{\alpha/2}(n)}$	$[\rho_-S^{*2},\rho_+S^{*2}]$
			$\chi^2=\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，$\mu$ 未知	$\rho_-=\dfrac{n-1}{\chi^2_{1-\alpha/2}(n-1)}$ $\rho_+=\dfrac{n-1}{\chi^2_{\alpha/2}(n-1)}$	$[\rho_-S^2,\rho_+S^2]$
$X_1\sim N(\mu_1,\sigma_1^2)$ $X_2\sim N(\mu_2,\sigma_2^2)$	$\mu_1-\mu_2$	$\bar{X}_1-\bar{X}_2$	$U=\dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$， σ_1^2,σ_2^2 已知	$\varepsilon=u_{1-\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$	$(\bar{X}_1-\bar{X}_2)\pm\varepsilon$
			$T=\dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $\sim t(n_1+n_2-2)$，$\sigma_1^2=\sigma_2^2$ 未知	$\varepsilon=t_{1-\alpha/2}(n_1+n_2-2)\cdot$ $S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$	$(\bar{X}_1-\bar{X}_2)\pm\varepsilon$

续表 3.2

<table>
<tr><th>背景</th><th>参数</th><th>估计量</th><th>枢轴量及其分布</th><th>误差限</th><th>置信区间</th></tr>
<tr><td rowspan="2">$X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2 \sim N(\mu_2, \sigma_2^2)$</td><td rowspan="2">$\frac{\sigma_1^2}{\sigma_2^2}$</td><td rowspan="2">$\frac{S_1^{*2}}{S_2^{*2}}$
或
$\frac{S_1^2}{S_2^2}$</td><td>$F=\frac{S_1^{*2}/\sigma_1^2}{S_2^{*2}/\sigma_2^2} \sim F(n_1, n_2)$,
μ_1, μ_2 已知</td><td>$\rho_- = \frac{1}{F_{1-\alpha/2}(n_1, n_2)}$
$\rho_+ = \frac{1}{F_{\alpha/2}(n_1, n_2)}$</td><td>$\left[\rho_- \frac{S_1^{*2}}{S_2^{*2}}, \rho_+ \frac{S_1^{*2}}{S_2^{*2}}\right]$</td></tr>
<tr><td>$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$,
μ_1, μ_2 未知</td><td>$\rho_- = \frac{1}{F_{1-\alpha/2}(n_1-1, n_2-1)}$
$\rho_+ = \frac{1}{F_{\alpha/2}(n_1-1, n_2-1)}$</td><td>$\left[\rho_- \frac{S_1^2}{S_2^2}, \rho_+ \frac{S_1^2}{S_2^2}\right]$</td></tr>
</table>

表 3.2 中，$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$，$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$，$S_w = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$.

【例 3.13】 从一批灯泡中随机抽取 5 只作寿命试验，测得寿命(单位：h)如下：1050，1100，1120，1250，1280. 设灯泡寿命服从正态分布. 试在 0.95 置信水平下估计灯泡的平均寿命.

分析　设 X 表示灯泡寿命，依题意 $X \sim N(\mu, \sigma^2)$，则灯泡的平均寿命为 $E(X) = \mu$. 因此本题的实质是估计正态分布参数 μ，但方差 σ^2 未知. 于是，参数 μ 的估计量选用样本均值 $\overline{X}$，枢轴量选用 $T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$. 而对寿命问题，通常只关心寿命下限，故相应的下侧区间估计的准则为 $P\{\mu \geqslant \hat{\mu}_L\} \geqslant 1 - \alpha$，其中置信下限 $\hat{\mu}_L = \overline{X} - t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$(注意：单侧估计时，显著性水平 α 不再等分配置在双侧尾部，而是全部置于所关注的一侧).

MATLAB 数据处理

```
clear
x = [1050,1100,1120,1250,1280];
N = length(x);
muEST = mean(x)
muLOWER = muEST - tinv(0.95,N - 1) * sqrt(var(x)/N)
```

上述指令的运行结果是：

```
muEST =
```

```
            1160
muLOWER =
            1.0649e+003
```

计算结果表明，这批灯泡的平均寿命约为 1160h，以 0.95 的概率保证这批灯泡的平均寿命不低于 1065h.

(2) 非正态变量分布参数的大样本估计方法

非正态变量分布参数的估计主要采用极大似然法，这是由于极大似然估计量优良的大样本性质. 设 $X \sim p(x;\theta)$，$\hat{\theta}_{MLE}$是分布参数 θ 的极大似然估计量，则在相当一般的条件下，下面两个结论成立.

① 强相合性：$P\{\lim\limits_{n\to\infty}\hat{\theta}_{MLE}=\theta\}=1$.

② 渐近正态性：$U=\sqrt{nI(\theta)}\,(\hat{\theta}_{MLE}-\theta)\xrightarrow{L}N(0,1)$，其中 $I(\theta)=-E_\theta\left(\frac{\partial^2}{\partial^2\theta}\ln p(x;\theta)\right)$ 是 Fisher 信息量，信息量 $I(\theta)$越大，变量分布中包含的关于未知参数 θ 的信息越多.

关于这两个结论的详细描述和证明参见文献[6].

当变量的分布非正态时，对分布参数 θ 进行估计的通常做法是：

① 求出参数 θ 的极大似然估计量$\hat{\theta}_{MLE}$；

② 根据渐近正态性，求出估计的边际误差 $\varepsilon=u_{1-\alpha/2}\sqrt{1/nI(\hat{\theta})}$；

③ 写出置信区间$[\hat{\theta}_{MLE}-\varepsilon,\hat{\theta}_{MLE}+\varepsilon]$.

【例 3.14】 设 $X \sim B(1,p)$，试估计分布参数 p.

解 因为变量 X 的概率分布律为

$$p(x;p)=p^x(1-p)^{1-x}\quad(0<p<1,x=0,1),$$

所以分布参数 p 的似然函数为

$$L(p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum\limits_{i=1}^{n}x_i}(1-p)^{n-\sum\limits_{i=1}^{n}x_i},$$

于是

$$\ln L(p)=\left(\sum_{i=1}^{n}x_i\right)\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p),$$

似然方程为

$$\frac{\mathrm{d}\ln L(p)}{\mathrm{d}p}=\frac{1}{p}\sum_{i=1}^{n}x_i-\frac{1}{1-p}\left(n-\sum_{i=1}^{n}x_i\right)=0,$$

解得

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

下面求 Fisher 信息量 $I(p)=-E\left(\frac{\partial^2}{\partial^2 p}\ln p(x;p)\right)$ 和边际误差 $\varepsilon=u_{1-\alpha/2}\cdot\sqrt{1/nI(\hat{p})}$.

因为

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln p(x;p)=\frac{\mathrm{d}}{\mathrm{d}p}[x\ln p+(1-x)\ln(1-p)]=\frac{x}{p}-\frac{1-x}{1-p},$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2 p}\ln p(x;p)=-\frac{x}{p^2}-\frac{1-x}{(1-p)^2},$$

所以

$$I(p)=-E\left(\frac{\mathrm{d}^2}{\mathrm{d}^2 p}\ln p(x;p)\right)=-E\left(-\frac{x}{p^2}-\frac{1-x}{(1-p)^2}\right)=\frac{E(x)}{p^2}+\frac{1-E(x)}{(1-p)^2}=\frac{1}{p(1-p)}.$$

于是边际误差

$$\varepsilon=u_{1-\alpha/2}\sqrt{1/nI(\hat{p})}=u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},$$

所以，由 $\hat{p}$ 可得参数 p 的 $1-\alpha$ 置信区间为

$$[\hat{p}-u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\hat{p}+u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}].$$

参数估计是一种重要的统计推断形式，这里介绍的基本上是由波兰统计学家 Neyman 所引进的方法. 由于对参数估计问题的不同理解，还有不同于 Neyman 方法的其他参数估计方法，如 Bayes 方法，限于篇幅这里没有加以介绍，有兴趣的读者请参阅文献[1]、[2]、[6].

关于区间估计的 MATLAB 数据处理，除例 3.13 依基本算法进行数据处理的做法之外，对于常用概率分布，可用 3.2.1 节中介绍的 mle 函数，只要选定返回第二个输出参数 pci，即可自动完成区间估计的工作.

【例 3.15】 引力常数的测定值 $X\sim N(\mu,\sigma^2)$，今分别使用金球和铂球进行实验测定.

(1) 用金球测定，观察值为：6.683, 6.681, 6.676, 6.678, 6.679, 6.672;

(2) 用铂球测定，观察值为：6.661, 6.661, 6.667, 6.667, 6.664.

试针对(1)、(2)两种情况分别对引力常数测定值的均值和标准差进行估计(置信水平为 0.9).

分析 此问题可依正态变量分布参数的小样本估计方法，对测定值均值的估计选估计量 $\overline{X}$ 和枢轴量 $T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$，置信区间为

$$\left[\overline{X}-t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \overline{X}+t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right].$$

对测定值标准差的估计选估计量 S^2 和枢轴量 $\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，置信区间为

$$\left[\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}S^2,\frac{n-1}{\chi^2_{\alpha/2}(n-1)}S^2\right].$$

然后，依上述算法组织 MATLAB 指令进行数据处理，这一工作留给读者练习. 这里，用 mle 函数进行数据处理.

MATLAB 数据处理(调用 mle 函数)

```
clear
x=[6.683    6.681    6.676    6.678    6.679    6.672];
y=[6.661    6.661    6.667    6.667    6.664];
[phat, pci] = mle(x, 'alpha', 0.1) % 金球测定的估计
[PHAT, PCI] = mle(y, 'alpha', 0.1) % 铂球测定的估计
```

上述指令的运行结果是：

```
phat =
      6.6782    0.0035
pci =
      6.6750    0.0026
      6.6813    0.0081
PHAT =
      6.6640    0.0027
PCI =
      6.6611    0.0019
      6.6669    0.0071
```

计算结果表明，金球测定的 μ 的估计值为 6.6782，置信区间为[6.6750, 6.6813]；σ 的估计值为 0.0035，置信区间为[0.0026, 0.0081]. 铂球测定的 μ 的估计值为 6.6640，置信区间为[6.6611, 6.6669]；σ 的估计值为 0.0027，置信区间为[0.0019, 0.0071].

除 mle 可用于参数的极大似然估计和区间估计之外，MATLAB 还给出了完成特定分布参数的极大似然估计和区间估计的 fit 类函数，内容见本书附录 B.

习题 3

1. 抽样调查某地区 50 户居民的月消费品支出额(单位：元)数据资料如下：

886	864	1027	918	866	926	893	919	946	978
928	1050	928	1040	905	900	900	863	926	821
999	927	978	854	954	999	800	981	895	924
946	949	816	1100	890	886	938	916	967	651
950	852	1000	900	1006	1120	864	818	921	850

试根据上述资料编制频率分布表和绘制频率直方图.

2. 设变量 X 服从区间$[0,\theta]$上的均匀分布，即分布密度为

$$p(x,\theta)=\begin{cases}\dfrac{1}{\theta}, & 0\leqslant x\leqslant\theta,\\ 0, & \text{其他}.\end{cases}$$

(1) 求参数 θ 的矩估计量 $\hat{\theta}_M$ 和 MLE $\hat{\theta}_L$；

(2) 现得样本值为 1.3, 0.6, 1.7, 2.2, 0.3, 1.1, 试分别用矩法与极大似然法求变量均值、变量方差的估计值.

3. 已知某种灯泡的寿命(单位：h)服从正态分布，在某周所生产的该种灯泡中随机抽取 10 只，测得其寿命为 1067, 919, 1196, 785, 1126, 936, 918, 1156, 920, 948. 设总体参数都为未知，试用极大似然法估计这周中生产的灯泡能使用 1300h 以上的概率.

4. 设变量 $\xi\sim N(\mu,\sigma^2)$，现得其样本值为 14.7, 15.1, 14.8, 15.0, 15.2, 14.6.

(1) 试用极大似然法与顺序统计量法估计变量的均值 μ；

(2) 试用极大似然法与顺序统计量法估计变量的方差 σ^2.

5. 随机地从一批钉子中抽取 16 枚，测得其长度(单位:cm)为 2.14, 2.10, 2.13, 2.15, 2.13, 2.12, 2.13, 2.10, 2.15, 2.12, 2.14, 2.10, 2.13, 2.11, 2.14, 2.11. 设钉长分布为正态的，试求总体均值 μ 的 90% 置信区间：

(1) 若已知 $\sigma=0.01$cm；

(2) 若 σ 为未知.

6. 某咨询公司调查了中国 20 个省级卫星电视频道晚 8 时至 9 时黄金档插播广告时间(单位：min)，假定总体服从正态分布. 统计数据如下:6.0, 6.6, 5.8, 7.0, 6.3, 6.2, 7.2, 5.7, 6.4, 7.0, 6.5, 6.2, 6.0, 6.5, 7.2, 7.3, 7.6, 6.8, 6.0, 6.2. 求中国省级卫星电视频道晚 8 时至 9 时黄金档插播广告时间均值的置信度为 95% 的置信区间.

7. 一项民意测验就某地区环境状况是否良好询问了 700 名成年人的看法，总共有 620 人的回答是“良好”.

(1) 求成年人中认为该地区环境状况良好的比率的点估计；

(2) 在 95% 的置信水平下，边际误差为多少?

(3) 求成年人中认为该地区环境状况良好的比率置信度为 90% 的置信区间.

8. 某高校有 3000 名走读生，该校拟估计这些学生每天往返学校的平均时间. 已知总体的标准差为 4.8min. 现要求进行置信度为 95%、抽样极限误差为 1min 的区间估计，试问按照重复抽样的方式，应抽取多大的样本?

9. 某减肥用品公司对其所做的报纸广告在两个城市的效果进行了比较，分别从两个城市中随机抽取了 800 名成年人，其中看过该广告的比例分别为 $p_1=19\%$，$p_2=16\%$，试求两城市中看过该广告的成年人比例之差的置信度为 95%的置信区间.

10. 随机地从甲批导线中抽取 4 根，乙批导线中抽取 5 根，测得电阻值(单位：Ω)如下：

甲：0.143, 0.142, 0.143, 0.137;

乙：0.140, 0.142, 0.136, 0.138, 0.140.

设甲、乙两批导线电阻分别服从 $N(\mu_1, 0.0025^2)$，$N(\mu_2, 0.0025^2)$，并且它们相互独立，但 μ_1，μ_2 未知，求 $\mu_1-\mu_2$ 的置信度为 0.95 的置信区间.

11. 某卷烟厂生产两种卷烟，现分别对两种卷烟的尼古丁含量做 6 次实验，结果如下.

甲：25, 28, 23, 26, 29, 22;

乙：28, 23, 30, 35, 21, 27.

若香烟的尼古丁含量服从正态分布，且方差相等，试求两种香烟的尼古丁平均含量差 $\mu_1-\mu_2$ 的 95%的置信区间.

12. 某自动机床加工同类型套筒，假设套筒的直径(单位：cm)服从正态分布. 现在从不同班次的产品中各抽取 5 个套筒，测定它们的直径数据如下.

A 班：2.066, 2.063, 2.068, 2.060, 2.068;

B 班：2.058, 2.057, 2.063, 2.059, 2.060.

试求两班所加工的套筒直径的方差之比 $\frac{\sigma_A^2}{\sigma_B^2}$ 的置信度为 0.90 的置信区间.

第 4 章 假设检验

假设检验是统计推断的另一个主要内容，它的基本任务是根据样本数据对变量是否服从某一特定分布或参数是否取某一特定的值等问题作出合理的判断.

本章讨论假设检验的基本问题，包括假设检验的基本概念、参数检验与分布拟合检验的常用方法.

4.1 假设检验概述

在统计应用中会遇到如下类型的问题.

【例 4.1】 一台自动车床在正常工作的情况下加工出的零件直径服从正态分布，零件规格是：标准直径 5cm，允许的最大加工误差 0.2cm. 某日开工后，技术人员进行例行检查，以判断该车床工作是否正常.

这是一个生产设备运行稳定性的监督问题. 在工业生产中监督设备的运行稳定性，通常的做法如下.

① 进行例行监督检查. 此时，往往假定设备的工作是正常的，然后每隔一段时间随机抽查几个产品的控制指标(如零件直径)，如果没有发现异常情况，就认为生产是正常的；如果发现产品的质量有大的变动，超过了允许的限度，则认为生产不正常而需要停机检修. 用统计语言描述就是，假设变量的分布形态已知，判断关于分布参数的一些已知信息是否为真，即进行变量分布参数的假设检验.

② 在生产环境发生变化，如设备大修或工艺改变等情况下，需要判断设备的运行是否符合正常状态要求，这不仅涉及①中所述的参数检验问题，首先要做的是判断产品的控制指标的概率分布是否与要求的一样. 用统计语言描述就是，对变量的分布形态已有先验的知识，如变量曾经或者应该服从正态分布、威布尔分布等，判断目前的情况是否果真如此.

假设检验是一类重要的、应用广泛的统计推断技术. 本章讨论假设检验的基本思想、方法和步骤等问题.

4.1.1 假设检验的思维逻辑

仍以例 4.1 中的问题为例，讨论假设检验的基本思想和方法. 假设这台自动车床的工作是正常的，零件直径服从正态分布，进行例行的质量检查. 假定从一天的产品中抽

查 50 个，分别测量直径，算得 $\overline{X}=4.8\text{cm}$. 据此来推断这台自动车床当天的生产是否正常.

这是变量分布参数的假设检验问题.

在假设检验问题的分析与推理中，首先要明确待检验的命题 H_0，称为**统计假设**(也叫**原假设**或**零假设**，称与之对立的假设 H_1 为**备择假设**)，然后由抽样结果来检查这个假设是否可信、是否能够成立，从而作出拒绝还是不拒绝这个假设的决策.

在例 4.1 中，一天中生产的所有零件的直径是一个随机变量 X，已知 X 服从正态分布．我们想知道，这一天生产的平均零件直径 $E(X)=\mu$ 是否符合标准要求，即 $\mu=5$ 是否成立. 如果 $\mu=5$，说明生产正常；否则，说明生产不正常.

于是，我们设原假设 $H_0: \mu=5$；备择假设 $H_1: \mu\neq5$.

怎样来判定 H_0 是否为真呢？由于 $X\sim N(\mu,\sigma^2)$，即 μ 是零件直径的期望值，而样本均值 $\overline{X}$ 是 μ 的性能优良的估计量，H_0 是否为真的判断可以通过定量分析二者的信息差异得到．现在 $\overline{X}=4.8$，而要求 $\mu=5$，其间存在差异 $\overline{X}-\mu=-0.2$，于是 H_0 是否为真取决于这个差异的性质.

① 差异可能是由随机因素引起的，称为**抽样误差**或**随机误差**，这种误差反映偶然的、非本质的因素引起的随机波动.

② 差异不是由随机因素引起的，它反映事物的本质差别(反映这天生产的平均零件直径同标准直径不同)，称为**系统误差**.

那么，这个抽样结果究竟是偶然性在起作用，还是该天生产不正常所造成的？这就需要给出一个量的界限．即给出一个小的正数 δ，如果 $|\overline{X}-\mu|<\delta$，则认为是随机性的差异，或者用统计学上的术语称**差异不够显著**；如果 $|\overline{X}-\mu|\geqslant\delta$，则认为不是随机性的差异，或者说**差异显著**.

于是，问题转化为如何确定这个正数 δ．容易想到，可以采用区间估计中的大概率置信准则

$$P\{|\overline{X}-\mu|<\delta\}\geqslant1-\alpha$$

来确定这个量的界限 δ.

但是，这里产生了一个问题：$\overline{X}$ 是一个随机变量，用 $\overline{X}$ 的观测值说明命题 $H_0: \mu=5$ 的真假是一种事实验证，若在一次抽样中 $|\overline{X}-\mu|<\delta$，只能增加人们对命题 H_0 的信心；即使是 100 次的验证都支持命题 H_0，但是仍不能令人信服命题 H_0 是真的.

如果注意到当 $X\sim N(\mu,\sigma^2)$时，有 $\overline{X}\sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$，即当 H_0 为真时，$\overline{X}$ 的观测值不应过于偏离 $\mu=5$，即事件 $\{|\overline{X}-\mu|\geqslant\delta\}$ 应当是一个小概率事件，不妨记为

$$P\{|\overline{X}-\mu|\geqslant\delta\}\leqslant\alpha,$$

称之为**检验准则**，其中 α 是一个很小的正数，称之为**显著性水平**．我们知道，**小概率事**

件在一次试验中基本上不会发生. 如果在一次抽样中, $\bar{X}$ 的样本观测值 $\bar{x}\in W$, 即 $\bar{X}$ 的观测值过于偏离 $\mu=5$, 试验结果与前提假设不相符, 则使人不能不怀疑作为这个小概率事件前提的命题 H_0 的正确性. 这里的集合 W 称为 H_0 的**拒绝域**. 如果一个概率很小的事件在一次试验中居然发生了, 则人们认为命题 H_0 不真的理由比承认命题 H_0 真更为充分. 也就是说, 在假设检验问题中, 采用伺机否定 H_0 的思维逻辑比执意支持 H_0 的思维逻辑更有说服力.

我们称在伺机否定 H_0 的思维过程中使用的推理方法为概率反证法, 它不同于一般的反证法. 一般的反证法如果在原假设下导出的结论自相矛盾或与事实矛盾, 则完全绝对地推翻原假设; 而概率反证法的结论不是绝对的, 只是认为结论正确的把握较大, 不排除犯错误的可能.

假设检验推理方法是概率反证法, 其推理逻辑是: 如果原假设 H_0 是对的, 而能够验证 H_0 为真的某个统计量落入某个约定的区域 W 是个小概率事件, 而小概率事件在一次试验中基本上不会发生. 如果该统计量的一次实测值落入区域 W, 也就是说, 原假设成立下的小概率事件在一次试验中发生了, 那么就以较充分的理由认为原假设不可信而否定它, 否则我们就不能否定原假设(只好接受它). 不否定原假设并不是肯定原假设一定对, 而只是说差异还不够显著, 还没有达到足以否定原假设的程度.

4.1.2 假设检验的基本步骤

假设检验的基本步骤如下.

第一步, 提出原假设 H_0 及备择假设 H_1.

原假设是我们对问题的标准统计描述, 是待验证的命题; 而备择假设则是原假设的对立命题, 是在否定原假设结论时的统计描述.

如例 4.1 中, 原假设 $H_0:\mu=\mu_0=5$; 备择假设 $H_1:\mu\neq\mu_0$.

我们称这类假设检验为**双侧假设检验**, 有时还会提出下述形式的假设:

$$H_0:\mu\leqslant\mu_0;\ H_1:\mu>\mu_0$$

或

$$H_0:\mu\geqslant\mu_0;\ H_1:\mu<\mu_0.$$

我们称这类假设检验为**单侧假设检验**.

此外要注意, 对于一个实际问题, 原假设通常都可以有两种提法, 即原假设和备择假设可以互换. 应该如何提取原假设呢? 这里给出一个原则性的建议: 在实际问题中, 往往把**系统久已存在或样本信息明显支持的状态、不宜轻易否定的命题**作为原假设 H_0, 或者说把我们**希望得到或反映系统新变化的结论**作为备择假设 H_1.

第二步, 选取一个适当的检验统计量 T, 并写出相应的检验准则.

如例 4.1 中，检验统计量为 $\bar{X}$，检验准则是 $P\{|\bar{X}-0.5|\geqslant\delta\}\leqslant\alpha$.

在这一环节应当注意，在 H_0 成立的条件下，所选定的检验统计量 T 的概率分布(或近似分布)应当是已知的．如例 4.1 中，若 H_0 成立，即 $X\sim N(5,0.2^2)$时，有 $\bar{X}\sim N(5,\ 0.0008)$.

拒绝域的临界值的计算依赖于检验统计量的概率分布．有时为了便于计算，特别是查表计算的情况下，需要对检验统计量进行分布形态规范化、标准化或渐近正态化变换．如例 4.1 中，通常需要将检验统计量 $\bar{X}$ 标准化变换为 $U=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$，在 $H_0:\mu=5$ 成立时 $U\sim N(0,1)$.

第三步，给定显著性水平 α，并求出 H_0 的拒绝域 W.

如例 4.1 中，给定的显著性水平 $\alpha=0.05$，由检验准则

$$P\{|\bar{X}-0.5|\geqslant\delta\}\leqslant\alpha,$$

可得

$$P\{\bar{X}\leqslant 0.5-\delta\}+P\{\bar{X}\geqslant 0.5+\delta\}\leqslant 0.05,$$

即

$$W=(-\infty,\ a]\cup[b,\ +\infty),$$

其中 $a=0.5-\delta$，$b=0.5+\delta$．通常用等分配置显著性水平的方法确定拒绝域的临界值，即

$$P\{\bar{X}\leqslant 0.5-\delta\}\leqslant 0.025,\quad P\{\bar{X}\geqslant 0.5+\delta\}\leqslant 0.025,$$

进而，根据 $\bar{X}\sim N(5,\ 0.0008)$，由 MATLAB 计算拒绝域的临界值.

```
a = norminv(0.025,5,0.0008)
b = norminv(0.975,5,0.0008)
```

上述指令的运行结果是：

```
a =
    4.9984
b =
    5.0016
```

即原假设 H_0 的拒绝域为 $W=(-\infty,\ 4.9984]\cup[5.0016,\ +\infty)$.

第四步，由样本算出检验统计量 T 的实测值，判断其是否落入拒绝域.

若实测值落入拒绝域，则认为差异显著而否定原假设 H_0；否则，就认为差异不显著而不能否定原假设，即保留(接受)原假设 H_0.

如例 4.1 中，$\bar{X}=4.8\in W$，故否定原假设 H_0，即认为这天生产不正常，需检修.

上面作出的否定原假设的判断，判断正确的可信程度为 0.95，判断错误的风险概率为 0.05.

4.1.3　检验的 p 值

在假设检验问题中，得出结论的依据是检验统计量 T 的观测值 t 是否落入原假设 H_0 的拒绝域 W. 如果 $t\in W$，则拒绝原假设 H_0，否则保留原假设 H_0. 这种非此即彼的结论有一个令人遗憾之处，即结论不能反映由当前的样本信息拒绝(或保留)原假设的理由是否充分. 具体地讲，统计量 T 的观测值 t 虽然落入拒绝域 W，但其距离 W 的临界值有多远? 如例 4.1 中，W 的左侧临界值为 4.998，检验统计量 $\bar{X}$ 的值为 4.8，小于 4.998，落入 W，我们拒绝原假设 H_0. 问题是：依据 $4.8<4.998$ 得出结论理由是否勉强? 对此最好有一个数量上的刻画. “检验的 p 值”能够满足人们的这种要求.

定义 4.1 (检验的 p 值)　设原假设为 H_0，T 是检验统计量，其观测值为 t，H_0 的拒绝域为 W，则称如下定义的概率 p 为原假设 H_0 的检验的 p 值.

若 $W=\{T: T\geqslant c\}$，则 $p=P(T\geqslant t\mid H_0\text{ 为真})$.

若 $W=\{T: T\leqslant c\}$，则 $p=P(T\leqslant t\mid H_0\text{ 为真})$.

若 $W=\{T: T\leqslant c_1\text{ 或 }T\geqslant c_2\}$，则

① 当 t 值较小(偏左取值)时，$p=2P(T\leqslant t\mid H_0\text{ 为真})$；

② 当 t 值较大(偏右取值)时，$p=2P(T\geqslant t\mid H_0\text{ 为真})$.

在统计实践中，人们并不事先指定显著性水平 α 的值，而是很方便地利用上面定义的 p 值. 对于任意大于 p 值的显著性水平，人们可以拒绝原假设，但不能在任何小于它的显著性水平下拒绝原假设. p 值是利用样本数据能够作出拒绝原假设的最小的显著性水平.

【例 4.2】　某人有 4 枚不同的硬币，他怀疑这 4 枚硬币的均匀性不同，想通过抛掷硬币观察出现正面的次数来鉴别硬币的均匀性. 于是进行了掷币试验，4 枚硬币各抛掷 100 次，并记录了出现正面的次数，结果见表 4.1.

表 4.1

硬币编号	1	2	3	4
出现正面的次数	50	55	60	65

分析　设在 100 次抛掷中每枚硬币出现正面的次数为 X_i，每次抛掷出现正面的概率分别为 $p_i(i=1,2,3,4)$，则 $X_i\sim b(100,\ p_i)$. 检验的原假设为

$H_0^{(i)}: p_i=p_0=0.5$(硬币是均匀的) $(i=1,2,3,4)$.

在 H_0 为真的假定下，即 $X_i\sim b(100,\ 0.5)$，出现正面的平均次数为 $E(X_i)=100\times 0.5=50$. 由于实测出现正面的次数均不小于 50，故可作单侧检验，即备择假设为

$$H_1^{(i)}: p_i>p_0=0.5\quad (i=1,2,3,4).$$

在显著性水平 α 下，检验准则是

$$P\{X_i-50\geqslant\delta\}\leqslant\alpha.$$

下面，我们利用 MATLAB 分别来求 H_0 的拒绝域和检验的 p 值.

MATLAB 数据处理

① 求拒绝域，这里指定显著性水平 $\alpha=0.05$. 由于检验统计量服从相同的分布，故对每种硬币原假设的拒绝域是相同的.

```
clear
Wlower = binoinv(0.95,100,0.5) % 求拒绝域的临界值 50+δ
```

上述指令的运行结果是：

```
Wlower =
       58
```

② 求对每种硬币进行检验的 p 值：$p_i=P\{X_i>x_i\}\quad(i=1,2,3,4)$.

```
clear
p1 = 1 - binocdf(50,100,0.5);
p2 = 1 - binocdf(55,100,0.5);
p3 = 1 - binocdf(60,100,0.5);
p4 = 1 - binocdf(65,100,0.5);
p = [p1,p2,p3,p4]
```

上述指令的运行结果是：

```
p =
   0.4602    0.1356    0.0176    0.0009
```

根据上述计算可知，在 0.05 显著性水平下，检验认为第 1 和第 2 两种硬币是均匀的，面第 3 和第 4 两种硬币不是均匀的.

如果改变显著性水平，则需重新计算拒绝域的临界值. 但是利用检验的 p 值进行决策则不必重新计算，应用起来更为灵活方便. 在 0.05 显著性水平下，检验的 p 值表明不必质疑第 1 种硬币均匀而第 4 种硬币不均匀的结论；如果严格均匀性的标准，即增大显著性水平(更容易拒绝原假设)，如取 0.15，则统计推断不能认为第 2 种硬币是均匀的；如果放宽均匀性的标准，即减小显著性水平(不容易拒绝原假设)，如取 0.01，则统计推断认为第 3 种硬币是均匀的.

4.1.4　假设检验中的两类错误与势函数

在假设检验方法的应用中，必须注意检验的结果是否与实际情况相吻合. 换句话说，假设检验是可能犯错误的. 在作出否定原假设的判断时，可能犯如下两类错误.

① 第一类错误. H_0 本来是正确的，但由于随机性使检验统计量的观测值落入拒绝域(小概率事件并非不可能发生)，依检验规则应当否定原假设. 这时的结论犯了“以真为假”的错误，即否定了正确的原假设.

显然，4.1.1 中讨论的检验准则是对检验中犯第一类错误的概率控制，即

$$P(\text{否定 } H_0 | H_0 \text{ 为真}) = P(\text{第一类错误}) = \alpha,$$

α 为事先给定的显著性水平.

② 第二类错误. 还有一种可能，如果原假设 H_0 是错误的，同样由于随机性使检验统计量的观测值没有落入拒绝域，依检验规则不能否定原假设. 这时的结论犯了"以假为真"的错误，即接受了错误的原假设. 犯第二类错误的概率记为

$$P(\text{不否定 } H_0 | H_0 \text{ 为假}) = P(\text{第二类错误}) = \beta,$$

或

$$P(\text{接受 } H_0 | H_1 \text{ 为真}) = P(\text{第二类错误}) = \beta.$$

我们希望检验的结论使犯两类错误的概率同时都很小，最好是全为 0. 但这是一个两难问题，当样本容量给定后，犯这两类错误的概率就不能同时被控制. 为了说明这种两难性，引入检验的势函数的概念.

定义 4.2（检验的势函数） 设 Θ 为 θ 的参数空间，$\Theta_0 \cup \Theta_1 = \Theta$ 且 $\Theta_0 \cap \Theta_1 = \varnothing$. 检验的原假设 $H_0: \theta \in \Theta_0$（备择假设为 $H_1: \theta \in \Theta_1$）的拒绝域为 W，则检验统计量 T 的观测值落入拒绝域 W 的概率

$$g(\theta) = P\{T \in W\} \quad (\theta \in \Theta)$$

称为该检验的势函数.

势函数实质上是对犯第一类错误的概率 $\alpha(=\alpha(\theta))$ 和犯第二类错误的概率 $\beta(=\beta(\theta))$ 的统一描述，是参数 θ 的函数，其关系式为

$$g(\theta) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0, \\ 1 - \beta(\theta), & \theta \in \Theta_1, \end{cases}$$

或

$$\alpha(\theta) = g(\theta) \quad (\theta \in \Theta_0),$$
$$\beta(\theta) = 1 - g(\theta) \quad (\theta \in \Theta_1).$$

为表述简单，在变量 $X \sim N(\mu, \sigma^2)$，σ^2 已知的条件下，以检验

$$H_0: \mu \geqslant \mu_0;\ H_1: \mu < \mu_0$$

为例对这一结论进行说明. 同例 4.1，这里 H_0 的检验统计量仍为 $\bar{X}$，拒绝域 $W = (-\infty, c]$，于是

$$g(\mu) = P\{\bar{X} \in W\} = P\{\bar{X} \leqslant c\} = P\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{c - \mu}{\sigma/\sqrt{n}}\right\} = \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right).$$

又由犯第一(二)类错误的概率 $\alpha(\beta)$ 的定义可知

当 $\mu \geqslant \mu_0$ 时，$g(\mu) = P\{\bar{X} \in W\} = P(\text{否定 } H_0 | H_0 \text{ 为真}) = \alpha$，即 α 是 μ 的函数；

当 $\mu<\mu_0$ 时，$g(\mu)=P\{\overline{X}\in W\}=P$(否定 $H_0|H_1$ 为真)$=1-P$(接受 $H_0|H_1$ 为真)$=1-\beta$，即 β 也是 μ 的函数.

显然，犯两类错误的概率可统一由势函数表示，即

$$\alpha(\mu)=g(\mu)=\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)\quad(\mu\geqslant\mu_0),$$

$$\beta(\mu)=1-g(\mu)=1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)\quad(\mu<\mu_0).$$

由这两个式子可以看出(σ 和 n 是确定的，$\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 是 c 的单调增函数)，欲使 α 减小，应使 $\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 中的 c 变小，此时导致 $1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 变大，即 β 变大；反之，欲使 β 减小，应使 $1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 变小，此时导致 c 变大，即 α 变大. 这就说明在假设检验的过程中，在给定样本容量的条件下，人们不可能使犯两类错误的概率同时都很小，即 α 与 β 之间一个变小必然导致另一个变大.

因此，在假设检验的实际应用时，通常人们只能控制犯第一类错误的概率，即根据实际情况，通过控制显著性水平 α 的大小来减少犯错误的可能性. 这种做法通常称为显著性检验.

在显著性检验过程中，当我们宁可“以假为真”而不愿“以真为假”时，则应把 α 取得很小，如 $\alpha=0.01$. 反之，则应把 α 取得大些，如 $\alpha=0.10$. 折中的取法是 $\alpha=0.05$. 例如，某药品含有毒性，必须严格控制不得超过规定的指标. 如果设原假设为产品不合格(毒性超过某一标准)，则应把 α 取得很小，这样才能保证用药的安全，当然难免会把一些合格品当成废品处理了. 在另一些情况下正好相反，例如检查袋装食品的质量，就没有必要那样严格，如果原假设为产品不合格(质量低于某标准)，可以把 α 取得稍大些，不管在什么情况下，为了保证 β 不致太大，样本容量都不应太小.

4.1.5 假设检验与区间估计的关系

假设检验与区间估计是两种最重要的统计推断形式. 这两者初看好像完全不同，其实两者之间有一定的联系. **利用区间估计可建立假设检验，反之亦然**. 下面仍用例 4.1 作简要说明.

设总体 $X\sim N(\mu,\sigma^2)$，σ^2 已知，若求 μ 的区间估计，应选择枢轴量

$$U=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),$$

按置信水平 $1-\alpha$ 确定一个大概率事件

$$P\left\{\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right|<u_{1-\alpha/2}\right\}=1-\alpha,$$

由此得到 μ 的置信水平为 $1-\alpha$ 的区间估计为

$$\left(\overline{X}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

这个区间估计恰好是原假设 $H_0:\mu=\mu_0$ 的一个接受区域，显著性水平为 α.

问题如果是检验假设

$$H_0:\mu=\mu_0;\quad H_1:\mu\neq\mu_0,$$

选取的统计量是

$$U=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1),$$

对给定的显著性水平 α，得到小概率事件

$$P\left\{\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}\right\}=\alpha,$$

由实测值 $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}$ 是否成立，决定是否拒绝原假设.

拒绝域为 $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}$，则接受域为 $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|<u_{1-\alpha/2}$，再把 μ_0 改为 μ，那么结果正是 μ 的区间估计，置信水平为 $1-\alpha$.

需要注意的是，假设检验和区间估计的结果在解释上是有差别的.

例如，我们在检验 $H_0:\mu=\mu_0=0$(显著性水平 α)的同时对 μ 作区间估计(置信水平为 $1-\alpha$)，可能会出现以下几种情况.

① **检验的结论与区间估计一致**. 如检验接受 H_0，区间估计为 $(-0.001,\ 0.001)$. 按假设检验，应接受 $\mu=0$；按区间估计，μ 可能取到的最大值和最小值都很接近 0，这两者解释一致.

② **区间估计强化了检验的结论**. 如检验拒绝 H_0，区间估计为(1000，2000). 按假设检验，应拒绝 $\mu=0$；按区间估计，区间中不包含 0，即 0 不看做 μ 的一个可能值，而且，区间的最小值也有 1000，与 0 相去甚远，故认为 $\mu\neq0$ 的理由很充分，区间估计的结论加强了假设检验的结论.

③ **检验的结论与区间估计不协调**. 如检验拒绝 H_0，区间估计为(0.001，0.002). 按假设检验，应拒绝 $\mu=0$；按区间估计，区间中不包含 0，从这个方面看两者一致. 可是细看这区间，就发现它整个在 0 的附近，因此实质上可以认为 μ 就是 0. 这样，区间估计的结论(在实质上)就与假设检验不同. 又如检验接受 H_0，区间估计为 $(-1000,\ 1500)$. 按假设检验，应接受 $\mu=0$；按区间估计，这区间包含 0，即 0 是 μ 的一个可能

值，在这一点上与假设检验的结论一致．但细看这区间，最大可以到 1500，最小可以到 -1000，这中间哪一个值都有可能．因此，从区间估计角度看，实在没有多大把握认为 μ 的取值都在 0 附近，这就与假设检验的结论不大协调了．

由此例可以看出，统计上的结论一定要注意其实质含义，如只停留在表面，就有可能被引入歧途．

4.2　变量分布参数的检验

4.2.1　正态变量均值与方差的假设检验

单正态变量均值与方差的假设检验的思维逻辑与步骤等同 4.1 节所述，例 4.1 就是单正态变量的均值检验问题．检验的关键是根据问题的特点，正确提出检验假设，选择恰当的检验统计量，然后根据检验统计量的概率分布求原假设拒绝域．下面给出正态变量均值和方差假设检验的方法要点，对方法的推导过程感兴趣的读者请参见文献[1]．

设变量 $X \sim N(\mu, \sigma^2)$，其样本均值为 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，样本方差为 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 或 $S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$，则正态变量均值和方差的假设检验法见表 4.2.

表 4.2　　正态变量均值和方差的假设检验法

参数	原假设	备择假设	检验条件	检验统计量及其分布	拒绝域
μ	$\mu \leqslant \mu_0$ $\mu \geqslant \mu_0$ $\mu = \mu_0$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	σ^2 已知	$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$U \geqslant u_{1-\alpha}$ $U \leqslant u_\alpha$ $\lvert U \rvert \geqslant u_{1-\alpha/2}$
			σ^2 未知	$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$	$t \geqslant t_{1-\alpha}(n-1)$ $t \leqslant t_\alpha(n-1)$ $\lvert t \rvert \geqslant t_{1-\alpha/2}(n-1)$
σ^2	$\sigma^2 \leqslant \sigma_0^2$ $\sigma^2 \geqslant \sigma_0^2$ $\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	μ 已知	$\chi^2 = \frac{nS^{*2}}{\sigma_0^2} \sim \chi^2(n)$	$\chi^2 \geqslant \chi^2_{1-\alpha}(n)$ $\chi^2 \leqslant \chi^2_\alpha(n)$ $\chi^2 \leqslant \chi^2_{\alpha/2}(n)$ 或 $\chi^2 \geqslant \chi^2_{1-\alpha/2}(n)$
			μ 未知	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2 \geqslant \chi^2_{1-\alpha}(n-1)$ $\chi^2 \leqslant \chi^2_\alpha(n-1)$ $\chi^2 \leqslant \chi^2_{\alpha/2}(n-1)$ 或 $\chi^2 \geqslant \chi^2_{1-\alpha/2}(n-1)$

下面举例说明上述检验法的应用．

【例4.3】 在正常生产情况下，印花棉布布幅的宽度服从正态分布 $N(1.4, 0.0048^2)$. 某日选取该种棉布5匹，测得布幅宽度为1.32，1.55，1.36，1.40，1.44，问该日印花棉布布幅宽度的标准差是否正常？(取 $\alpha=0.05$)

分析 这是正态分布的方差检验问题. 依题意，令 $H_0:\sigma=0.0048; H_1:\sigma\neq 0.0048$. 检验统计量选样本方差 S^2，双侧检验，检验准则为

$$P(S^2\leqslant\delta_1|\sigma=0.0048)\leqslant\alpha/2 \quad 或 \quad P(S^2\geqslant\delta_2|\sigma=0.0048)\leqslant\alpha/2,$$

需对 S^2 进行变换以确定其概率分布. 由抽样分布理论 $\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$，故在 H_0 成立的条件下，$\chi^2=4S^2/0.0048^2\sim\chi^2(4)$，即

$$P\{\chi^2\leqslant 4\delta_1/0.0048^2\}\leqslant\alpha/2 \quad 或 \quad P\{\chi^2\geqslant 4\delta_2/0.0048^2\}\leqslant\alpha/2,$$

由此可求出 H_0 拒绝域的临界值.

MATLAB 数据处理

```
clear
x=[1.32,1.55,1.36,1.40,1.44];
XVAR=var(x) % 求检验统计量的值
DETA1=chi2inv(0.025,4)*0.0048^2/4 % 求拒绝域的左侧临界值
DETA2=chi2inv(0.975,4)*0.0048^2/4 % 求拒绝域的右侧临界值
p=1-chi2cdf(4*XVAR/0.0048^2,4) % 求检验的 p 值
```

上述指令的运行结果是：

```
XVAR =
    0.0078
DETA1 =
    2.7903e-006
DETA2 =
    6.4185e-005
p =
    0
```

由于检验统计量实测值 $S^2=0.0078>$ DETA2 $=0.000064185$，落入拒绝域，故否定原假设，即认为该日生产棉布布幅宽度的标准差不正常；检验的 p 值近似为零，表明作出这一结论的理由是充分的.

MATLAB统计工具箱给出了两个用于正态分布均值检验的函数，它们是方差已知条件下的 U 检验法函数 ztest，和方差未知条件下的 t 检验法函数 ttest. 下面举例说明这两个函数的使用.

【例 4.4】 某车间用一台包装机包装葡萄糖，包得的袋装糖重是一个随机变量，它服从正态分布．当机器正常时，其均值为 0.5kg，标准差为 0.015kg．某日开工后检验包装机是否正常，随机地抽取所包装的糖 9 袋，称得净重(单位：kg)为

0.497，0.506，0.518，0.524，0.498，0.511，0.52，0.515，0.512，

问机器是否正常？

分析 这是方差已知条件下正态分布均值的检验问题．注意到多数样本数据大于 0.5，故作单侧检验，检验假设为 H_0：$\mu=\mu_0=0.5$；H_1：$\mu>0.5$．注意，这里的原假设与备择假设是不相容的，但并非完全对立．这也是在实际应用中经常采用的检验命题的设定技巧．

MATLAB 数据处理

调用 U 检验法函数 ztest 函数，其调用格式为

```
[h, p, ci, U] = ztest(x, m, sigma, alpha, tail)
```

其中，输入参数 x 为样本数据向量，m 为待检验均值，sigma 为正态分布的标准差，alpha 为显著性水平(默认值 0.05)，tail 为检验的备择假设的标示值(tail=0 表示双侧检验，tail=1 表示右侧检验“>”，tail= −1 表示左侧检验“<”)；输出参数 h 为检验决策值(h=0 表示在显著性水平 alpha 下不能拒绝原假设，h=1 表示在显著性水平 alpha 下可以拒绝原假设)，p 为拒绝原假设的最小显著性概率，ci 为真实均值 μ 的 1－alpha 置信区间，U 为检验统计量的值．

```
clear
x=[0.497,0.506,0.518,0.524,0.498,0.511,0.52,0.515,0.512];
[h,p,ci,U]=ztest(x,0.5,0.015,0.05,1)
```

上述指令的运行结果是：

```
h =
     1
p =
     0.0124
ci =
     0.5030        Inf
U =
     2.2444
```

结果表明在 0.05 显著性水平下，可拒绝原假设，即认为包装机工作不正常，每袋葡萄糖的平均质量大于 0.5kg，由 ci 的值可知每袋葡萄糖的平均质量不低于 0.503kg 的可信程度为 0.95．

若忽视每袋葡萄糖质量的标准差已知的条件，则可调用函数 ttest 完成检验工作，其

调用格式同 ztest.

```
[h,p,ci,T] = ttest(x,0.5,0.05,1)
```

上述指令的运行结果是：

```
h =

     1

p =

     0.0036

ci =

     0.5054        Inf

T =

      tstat:  3.5849
         df:  8
         sd:  0.0094
```

结果表明在 0.05 显著性水平下，t 检验亦拒绝原假设，即认为包装机工作不正常，每袋葡萄糖的平均质量大于 0.5kg；且由 p 值可知，这个结论在 0.01 显著性水平下也是站得住脚的. 由 ci 的值可知每袋葡萄糖的平均质量不低于 0.5054kg 的可信程度为 0.99，结论错误的风险概率是 0.01. 输出参数 T 报告检验统计量的观测值 tstat = 3.5849，t 分布的自由度 df = 8，对每袋葡萄糖质量标准差的估计 sd = 0.0094.

这里对例 4.4 稍作引申. 生产商为确保产品投放市场后不出现较多的因质量指标不合格而引起的消费者投诉，在生产过程中实际的装袋质量往往大于向市场承诺的标准质量. 在此例中，如果我们将袋装葡萄糖的平均质量 0.5kg、标准差 0.015kg 理解成是生产商对产品质量指标的承诺(而不是包装机的实际生产控制指标)，则由每袋葡萄糖质量的样本标准差小于 0.01kg(更小于 0.015kg)可以认为，包装机的工作状态是平稳的. 因此，样本均值大于 0.5kg 应是生产商确保质量指标承诺的体现. 实际上，若以样本均值和样本标准差作为包装机的实际控制参数(估计)，则可以推算出该生产商投放到市场上的袋装葡萄糖每袋质量大于 0.5kg 的比率，如下所示.

```
p = 1 - normcdf(0.5,mean(x),std(x))
```

上述指令的运行结果是：

```
p =

     0.8840
```

即 88% 的袋装葡萄糖的质量大于 0.5kg.

4.2.2 两个正态变量均值与方差的比较

两个正态变量均值和方差的比较，等价于两个正态变量均值差和方差比的假设检

验. 检验的思维逻辑与步骤同前所述一致, 问题的关键是正确提出检验假设, 选择恰当的检验统计量, 然后根据检验统计量的概率分布求出原假设的拒绝域. 下面给出两个正态变量均值差和方差比检验的方法要点, 对方法的推导过程感兴趣的读者参见文献[1].

设变量 $X\sim N(\mu_1,\ \sigma_1^2)$, 变量 $Y\sim N(\mu_2,\ \sigma_2^2)$, 样本均值分别为 $\overline{X}$ 和 $\overline{Y}$, 样本方差分别为 S_X^2 和 S_Y^2(或同 4.2.1 节, 为 S_X^{*2} 和 S_Y^{*2}), 记

$$S_w^2=\frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{n_1+n_2-2},\quad l=\left(\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}\right)^2\Big/\left(\frac{S_X^4}{n_1^2(n_1-1)}+\frac{S_Y^4}{n_2^2(n_2-1)}\right),$$

则两个正态变量均值差和方差比的假设检验法见表 4.3.

表 4.3　　两个正态变量均值差和方差比的假设检验法

参数	原假设	备择假设	检验条件	检验统计量	拒绝域
$\mu_1-\mu_2$	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	σ_1,σ_2 已知	$U=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$U\geqslant u_{1-\alpha}$ $U\leqslant u_\alpha$ $\|U\|\geqslant u_{1-\alpha/2}$
			$\sigma_1=\sigma_2=\sigma$ 未知	$t=\dfrac{\overline{X}-\overline{Y}}{S_w\Big/\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$	$t\geqslant t_{1-\alpha}(n_1+n_2-2)$ $t\leqslant t_\alpha(n_1+n_2-2)$ $\|t\|\geqslant t_{1-\alpha/2}(n_1+n_2-2)$
			σ_1,σ_2 未知 n_1,n_2 充分大	$U^*=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_X^2}{n_1}+\dfrac{S_Y^2}{n_2}}}$	$U^*\geqslant u_{1-\alpha}$ $U^*\leqslant u_\alpha$ $\|U^*\|\geqslant u_{1-\alpha/2}$
			σ_1,σ_2 未知 n_1,n_2 不够充分大	$t^*=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_X^2}{n_1}+\dfrac{S_Y^2}{n_2}}}$	$t^*\geqslant t_{1-\alpha}(l)$ $t^*\leqslant t_\alpha(l)$ $\|t^*\|\geqslant t_{1-\alpha/2}(l)$
$\dfrac{\sigma_1}{\sigma_2}$	$\sigma_1^2\leqslant\sigma_2^2$ $\sigma_1^2\geqslant\sigma_2^2$ $\sigma_1^2=\sigma_2^2$	$\sigma_1^2>\sigma_2^2$ $\sigma_1^2<\sigma_2^2$ $\sigma_1^2\neq\sigma_2^2$	μ_1,μ_2 已知	$F^*=\dfrac{S_X^{*2}}{S_Y^{*2}}$	$F^*\geqslant F_{1-\alpha}(n_1,n_2)$ $F^*\leqslant F_\alpha(n_1,n_2)$ $F^*\leqslant F_{\alpha/2}(n_1,n_2)$或$F^*\geqslant F_{1-\alpha/2}(n_1,n_2)$
			μ_1,μ_2 未知	$F=\dfrac{S_X^2}{S_Y^2}$	$F\geqslant F_{1-\alpha}(n_1-1,n_2-1)$ $F\leqslant F_\alpha(n_1-1,n_2-1)$ $F\leqslant F_{\alpha/2}(n_1-1,n_2-1)$或$F\geqslant F_{1-\alpha/2}(n_1-1,n_2-1)$

表 4.3 中, 检验统计量的概率分布为: $U\sim N(0,1)$, $t\sim t(n_1+n_2-2)$, U^* 近似服从标准正态分布, t^* 近似服从自由度为 l 的 t 分布, $F^*\sim F(n_1,n_2)$, $F\sim F(n_1-1,\ n_2-1)$.

下面举几个例子, 以巩固对上述检验法的理解.

【例4.5】 设甲、乙两煤矿出煤的含灰率(单位：%)都服从正态分布，即 $X \sim N(\mu_1, 7.5)$，$Y \sim N(\mu_2, 2.6)$，为检验两煤矿的煤含灰率有无显著性差异，从两矿中各取样若干份，分析结果如下.

甲矿：24.3，20.8，23.7，21.3，17.4；

乙矿：18.2，16.9，20.2，16.7.

试在显著性水平 $\alpha = 0.05$ 下，检验"含灰率无差异"这个假设.

分析 检验假设为

$$H_0: \mu_1 = \mu_2; \quad H_1: \mu_1 \neq \mu_2.$$

取检验统计量 $\bar{X} - \bar{Y}$，由于 σ_1^2，σ_2^2 均已知，统计量规范化为 $U = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$，

检验准则是 $P\{|U| \geqslant \delta\} \leqslant \alpha$，即拒绝域为 $|U| \geqslant \delta$.

MATLAB 数据处理

```
clear
x = [24.3, 20.8, 23.7, 21.3, 17.4];
y = [18.2, 16.9, 20.2, 16.7];
alpha = 0.05; % 设定显著性水平
U = (mean(x) - mean(y))/sqrt(7.5/5 + 2.6/4); % 计算检验统计量的观测值
DETA = norminv((1 - alpha/2), 0, 1); % 求拒绝域的临界值
p = 1 - normcdf(U, 0, 1); % 求拒绝原假设的最小显著性概率
if abs(U) > DETA % 决策，拒绝原假设则返回 h=1，否则返回 h=0
h = 1;
else
h = 0;
end
alpha, h, p, U, DETA
```

上述指令的运行结果是：

```
alpha =
      0.0500
h =
      1
p =
      0.0085
```

```
U =
      2.3870
DETA =
      1.9600
```

结果表明在 0.05 的显著性水平下，认为甲矿含灰率与乙矿含灰率有显著差异.

若注意到含灰率数据的均值甲矿明显大于乙矿，进行单侧检验更为恰当，检验假设可表示为

$$H_0:\mu_1=\mu_2;\quad H_1:\mu_1>\mu_2.$$

此时，检验准则是 $P\{U\geqslant\delta\}\leqslant\alpha$，即拒绝域为 $U\geqslant\delta$. 相应的数据处理过程只需在上述 MATLAB 指令集中，将语句

```
DETA = norminv((1 - alpha/2),0,1)
```

修改为

```
DETA = norminv((1 - alpha),0,1)
```

即可. 此时 DETA =1.6449，其他计算结果不变. 相应的检验结论是：在 0.05 的显著性水平下，认为甲矿含灰率显著地大于乙矿含灰率. 由 p 值可知，这个结论在 0.01 的显著性水平下也是成立的.

MATLAB 给出了方差未知但等方差条件下用于两个正态变量均值差的检验函数 ttest2，使用方法与 ttest 类似.

【例 4.6】 在平炉上进行一项试验以确定改变操作方法的建议是否会增加钢的产率（单位：%），试验是在同一只平炉上进行的. 每炼一炉钢时除操作方法外，其他条件都尽可能做到相同. 先用标准方法炼一炉，然后用建议的新方法炼一炉，以后交替进行，各炼 10 炉，其产率分别如下.

① 标准方法：78.1，72.4，76.2，74.3，77.4，78.4，76.0，75.5，76.7，77.3；

② 新 方 法：79.1，81.0，77.3，79.1，80.0，79.1，79.1，77.3，80.2，82.1.

设这两个样本相互独立，并且钢的产率服从正态分布. 问建议的新操作方法能否提高产率？（取 $\alpha=0.05$）

分析 这是两个正态变量均值的比较问题，应作均值差的检验. 由于变量的方差未知且样本容量较小，故应在等方差的假定下进行 t 检验. 因此，此问题严谨的分析应当分如下两步.

① 作方差齐性检验，即检验 $H_0:\sigma_1^2=\sigma_2^2$；$H_1:\sigma_1^2\neq\sigma_2^2$.

② 方差齐性检验通过的情况下作均值差 t 检验(若等方差的假定不成立，则只能作近似 t 检验)，即检验 $H_0:\mu_1=\mu_2$；$H_1:\mu_1<\mu_2$.

MATLAB 数据处理

① 方差齐性检验，取检验统计量 $F=\frac{S_1^2}{S_2^2}\sim F(9,9)$，$H_0$ 的拒绝域为 $F\leqslant F_{0.025}(9,9)$ 或 $F\geqslant F_{0.975}(9,9)$.

```
clear
x = [78.1,72.4,76.2,74.3,77.4,78.4,76.0,75.5,76.7,77.3];
y = [79.1,81.0,77.3,79.1,80.0,79.1,79.1,77.3,80.2,82.1];
F = var(x)/var(y);
p = 1 - fcdf(F,9,9)
```

上述指令的运行结果是：

```
p =
   0.2795
```

结果表明，可以拒绝 H_0 的最小显著性概率 $p=0.2795>\alpha=0.05$，故不能拒绝 H_0，即认为标准方法与新方法钢的产率方差是一致的，这也说明试验中除操作方法外，其他条件都得到了较好的控制.

② 均值差 t 检验，调用函数 ttest2.

```
[h,p,ci,TT] = ttest2(x,y,0.05,-1)
```

上述指令的运行结果是：

```
h =
   1
p =
   2.1759e-004
ci =
   -Inf   -1.9083
TT =
   tstat: -4.2957
      df: 18
      sd: 1.6657
```

结果表明，可以拒绝 H_0，即新操作方法能显著提高钢的产率. 由 p 值可知结论错误的可能性极低(小于 1%)，由 ci 的上限值可知 $\mu_2-\mu_1>1.9$，即有 99%以上的把握新方法能提高钢的产率(经计算)约 2.5 个百分点，实际生产中钢的产率在 ±1.67 范围内波动.

4.2.3 非正态变量分布参数的检验

关于非正态变量分布参数的检验，除少数特殊分布可在小样本条件下进行检验之

外，通常都是在大样本条件下进行近似检验.

4.2.3.1 几种特殊分布参数的小样本检验

(1) 0－1分布参数 p 的检验

0－1分布参数 p 的检验，是最重要的、应用广泛的非正态分布参数的检验问题，人们习惯上称为比率 p 的检验.

下面，结合实例来阐述比率 p 的检验方法.

【例4.7】 某机床加工的零件长期以来不合格率不超过0.01，某天开工后，为检验机床工作是否稳定，随机抽检了15件产品，发现其中有1件不合格，试问该机床是否需要检修.

设 X 为抽检出的一件产品的不合格数，则 X 服从0－1分布 $b(1,p)$，其中 p 为产品的不合格率，$0<p<1$. 当机床工作稳定时 $p\leqslant 0.01$，当机床工作不稳定时 $p>0.01$. 因此，判断该机床是否需要检修的问题可由如下假设检验问题作出推断：

$$H_0: p\leqslant 0.01;\quad H_1: p>0.01.$$

这是一个离散分布的单边检验问题. 设 $X_1, X_2, \cdots, X_n$ i.i.d.$\sim X$，由于 $E(X)=p$，所以选取 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 为检验统计量，在 n 确定时可以用 $T=\sum_{i=1}^{n}X_i$.

当 H_0 为真时，$\bar{X}$ 不应过大，即 T 不会过大；反之，当 H_0 不真时，$\bar{X}$ 较大，即 T 会取较大的值. 因此，H_0 的拒绝域的形式为 $W=\{T\geqslant c\}$，这里 c 是临界值. 问题的关键是如何求得临界值 c.

当 $p=p_0$ 时，统计量 $T\sim b(n,p_0)$，故可用二项分布来决定临界值 c. 由于 T 取非负整数，故 c 亦应取非负整数.

给定显著性水平 α，检验准则为 $P(T\geqslant c\mid p=p_0)\leqslant\alpha$，此时拒绝域 W 的大小受到限制(即存在 c_0，当 $c=c_0$ 时，拒绝域 W 不能再扩大). 于是，临界值 c 可取满足

$$P_{p_0}\{T\geqslant c\}=\sum_{i=c}^{n}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha$$

的最小整数.

同理可以得出比率 p 检验的其他两种情形的检验方法. 下面对比率 p 的检验问题作一般叙述.

设 $X_1, X_2, \cdots, X_n$ i.i.d.$\sim b(1,p)$，则关于参数 p 的检验问题与方法见表4.4.

根据上述讨论，下面给出例4.7的具体检验过程. 在例4.7中，当 H_0 为真时，检验统计量 $T=\sum_{i=1}^{n}X_i\sim b(15,0.01)$，拒绝域为 $W=\{T\geqslant c\}$，临界值 c 是满足

$$\sum_{i=c}^{n}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha$$

的最小整数. 检验的 MATLAB 数据处理如下.

表4.4 0-1分布参数 p 的假设检验法(显著性水平为 α, 检验统计量 $T=\sum X_i$)

原假设	备择假设	拒绝域	临界值 c 的确定方法
$p\leqslant p_0$	$p>p_0$	$T\geqslant c$	c 取满足 $\sum_{i=c}^{n}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha$ 的最小整数
$p\geqslant p_0$	$p<p_0$	$T\leqslant c$	c 取满足 $\sum_{i=0}^{c}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha$ 的最大整数
$p=p_0$	$p\neq p_0$	$T\leqslant c_1$ 或 $T\geqslant c_2$, $c_1<c_2$	c_1 取满足 $\sum_{i=0}^{c_1}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha/2$ 的最大整数 c_2 取满足 $\sum_{i=c_2}^{n}\binom{n}{i}p_0^i(1-p_0)^{n-i}\leqslant\alpha/2$ 的最小整数

```
clear
T=1; %检验统计量的观测值
alpha=0.05; %显著性水平
p=1-binocdf(0:15, 15, 0.01); %为确定拒绝域临界值计算 T≥c 的概率
for byk=1:16 %求拒绝域临界值
if p(byk)>alpha&p(byk+1)<=alpha
c=byk;
end
end
if T>=c %检验决策, h=1(0)拒绝(接受)原假设
h=1
else
h=0
end
```

上述指令的运行结果是:

```
h =
    1
```

结果表明, 拒绝原假设, 即统计推断认为应检修机床.

【例4.8】 某厂产品的优质品率一直保持在40%, 近期技监部门来厂抽查, 共抽查了12件产品, 其中优质品为5件, 在0.05显著性水平下能否认为其优质品率仍保持在40%?

分析 设 X 表示检查一个产品时优质品的个数, 则 $X\sim b(1,p)$. 检验问题为

$$H_0: p=0.4; \quad H_1: p\neq 0.4.$$

这是一个双边检验问题. 当 H_0 为真时, 检验统计量 $T=\sum_{i=1}^{n}X_i \sim b(12,\ 0.4)$, 拒绝域为 $T\leqslant c_1$ 或 $T\geqslant c_2(c_1<c_2)$. 其中, 临界值 c_1 是使 $P\{T\leqslant c_1\}\leqslant 0.025$ 成立的最大整数, c_2 是使 $P\{T\geqslant c_2\}\leqslant 0.025$ 成立的最小整数.

MATLAB 数据处理

```
clear
T=5; %检验统计量的观测值
alpha=0.025; %显著性水平
p=binocdf(0:12,12,0.4); %为确定拒绝域临界值计算T的累积概率
for byk=1:7 %求拒绝域临界值
if p(byk)<alpha&p(byk+1)>=alpha
c1=byk-1;
end
if (1-p(byk+6))>alpha&(1-p(byk+7))<=alpha
c2=byk+7;
end
end
if T<=c1|T>=c2 %检验决策, h=1(0)拒绝(接受)原假设
h=1
else
h=0
end
c=[c1,c2] %输出拒绝域临界值
```

上述指令的运行结果是:

```
h =
      0
c =
      1      9
```

上述计算表明, 当 $\alpha=0.05$ 时, 由于 $P\{T\leqslant 1\}<0.025$ 而 $P\{T\leqslant 2\}>0.025$, 故拒绝域左侧临界值 $c_1=1$; 又 $P\{T\geqslant 8\}>0.025$ 而 $P\{T\geqslant 9\}<0.025$, 故拒绝域右侧临界值 $c_2=9$. 于是, H_0 的拒绝域为 $T\leqslant 1$ 或 $T\geqslant 9$. 检验统计量的观测值 $T=5$ 未落入拒绝域, 因而在 0.05 显著性水平下认为该厂优质品率无明显变化.

(2) 泊松分布参数 λ 的检验

泊松分布在描述稀有事件发生次数方面发挥着重要的作用. 下面结合实例来阐述泊松分布参数 λ 的检验方法.

【例 4.9】 通常认为放射性物质在单位时间内放射的 α 粒子数 X 服从泊松分布 $P(\lambda)$. 其中 λ 是单位时间内平均放射的 α 粒子数. 要测试某放射性污染地区的单位时间内平均放射的 α 粒子数是否超过临界值 λ_0.

这是泊松分布参数 λ 的检验问题, 所要检验的假设是

$$H_0:\lambda\leqslant\lambda_0;\quad H_1:\lambda>\lambda_0.$$

设在 n 个单位时间内测得的 α 粒子数 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$. 由于 $E(X)=\lambda$, 因此选择 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 为检验统计量, 在 n 确定时可以用 $T=\sum_{i=1}^{n}X_i$. 很显然, T 值越大越对 H_0 不利, 因此 H_0 的拒绝域应具有 $T\geqslant c$ 的形式. 由泊松分布的可加性, $T\sim P(n\lambda)$, 所以检验准则为 $P(T\geqslant c\mid\lambda=\lambda_0)\leqslant\alpha$(显著性水平), 即拒绝域的临界值 c 应是满足

$$P_{\lambda_0}\{T\geqslant c\}=\sum_{k=c}^{\infty}\frac{(n\lambda_0)^k}{k!}\mathrm{e}^{-n\lambda_0}\leqslant\alpha$$

的最小正整数. 在实际计算中, 常常利用泊松分布与 χ^2 分布的如下关系:

对给定的 λ 及 $T\sim P(n\lambda)$, 有

$$P\{T\geqslant c\}=\sum_{k=c}^{\infty}\frac{(n\lambda)^k}{k!}\mathrm{e}^{-n\lambda}=\chi^2(2n\lambda;\ 2c).$$

其中, $\chi^2(2n\lambda;\ 2c)$ 表示自由度为 $2c$ 的 χ^2 分布在 $2n\lambda$ 处的值. 显然, $P\{T\geqslant c\}$ 是 λ 的单调增函数.

对这个结论的证明感兴趣的读者可参考泊松分布与伽玛分布关系的讨论, χ^2 分布是一种特殊的伽玛分布, 初步的讨论参见文献[1].

于是, 拒绝域的临界值 c 应是满足 $\chi^2(2n\lambda_0;\ 2c)\leqslant\alpha$ 即 $2n\lambda_0\leqslant\chi^2_{1-\alpha}(2c)$ 的最小正整数.

同理可得其余两种检验问题的检验方法, 结果列于表 4.5.

表 4.5　泊松分布参数 λ 的假设检验法(显著性水平为 α, 检验统计量 $T=\sum X_i$)

原假设	备择假设	拒绝域	临界值 c 的确定方法
$\lambda\leqslant\lambda_0$	$\lambda>\lambda_0$	$T\geqslant c$	c 取满足 $2n\lambda_0\leqslant\chi^2_{1-\alpha}(2c)$ 的最小整数
$\lambda\geqslant\lambda_0$	$\lambda<\lambda_0$	$T\leqslant c$	c 取满足 $2n\lambda_0\leqslant\chi^2_{\alpha}(2c+2)$ 的最大整数
$\lambda=\lambda_0$	$\lambda\neq\lambda_0$	$T\leqslant c_1$ 或 $T\geqslant c_2, c_1<c_2$	c_1 取满足 $2n\lambda_0\geqslant\chi^2_{\alpha/2}(2c_1+2)$ 的最大整数 c_2 取满足 $2n\lambda_0\leqslant\chi^2_{1-\alpha/2}(2c_2)$ 的最小整数

继续讨论例 4.9. 通常，单位时间(时间长度为 90min)内平均放射的 α 粒子数不超过 0.6. 假定进行了 15 次观测，观测到的 α 粒子数见表 4.6.

表 4.6

粒子数 a_i	0	1	2	3	4	合计
频数 n_i	4	7	2	1	1	15

于是，检验原假设为 $H_0:\lambda\leqslant 0.6$，取显著性水平为 0.1. 由上述讨论，检验的 MATLAB 数据处理如下.

```
clear
A=[0,1,2,3,4]; % 粒子数数据
N=[4,7,2,1,1]; % 频数数据
T=N*A' % 检验统计量的观测值
alpha=0.1; % 显著性水平
n=sum(N); % 样本容量
lambda0=0.6; % 待检验参数值
c=0.5*chi2inv(1-alpha, 2*n*lambda0) % 求拒绝域临界值
if T>=c % 检验决策, h=1(0)拒绝(接受)原假设
h=1
else
h=0
end
```

上述指令的运行结果是:

```
T =
     18
c =
     12.9947
h =
     1
```

结果表明，T<c，h=1，拒绝原假设，即放射性污染地区的单位时间内平均放射的 α 粒子数超过临界值 0.6.

(3) 指数分布参数 θ 的检验

指数分布是一类重要的分布，应用广泛. 下面结合实例来阐述指数分布参数 θ 的检验方法.

【例 4.10】 设一批电子元件，其寿命 X(单位：h)服从参数为 θ 的指数分布. 假定

从这批元件中随机抽取 n 个样品，进行加速寿命试验，并测得全部 n 个样品的失效时间. 假定按照国家标准，这种电子元件的平均寿命不得低于 θ_0 h. 又假定在加速寿命试验中样品的平均寿命为正常状态下的$\frac{1}{10}$. 如何根据上述信息判定这批电子元件是否合乎标准?

根据上述信息判定这批电子元件是否合乎标准的问题，等价于指数分布参数 θ 的检验问题

$$H_0:\theta\geqslant\theta_0;\quad H_1:\theta<\theta_0.$$

设 n 个样品在正常情况下的失效时间 $X_1,X_2,\cdots,X_n$ i.i.d$\sim X$. 由于 $E(X)=\theta$，因此选择 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 为检验统计量. 很显然，$\bar{X}$ 值越小越对H_0 不利，因此 H_0 的拒绝域应具有 $\bar{X}\leqslant c$ 的形式.

又由指数分布是特殊的伽玛分布，即 $\mathrm{Exp}(1/\theta)=\mathrm{Ga}(1,\ 1/\theta)$，$n$ 个独立同分布指数变量之和为伽玛变量可知，$n\bar{X}=\sum_{i=1}^{n}X_i\sim\mathrm{Ga}(n,\ 1/\theta)$. 为计算简便，通常利用伽玛分布的性质引进一个 χ^2 统计量作为检验统计量，在 $\theta=\theta_0$ 时，$\chi^2=2n\bar{X}/\theta_0\sim\chi^2(2n)$.

于是，在显著性水平 α 下，由检验准则 $P(\bar{X}\leqslant c\mid\theta=\theta_0)\leqslant\alpha$ 可知，H_0 的拒绝域取 $\bar{X}\leqslant c$ 与取$\chi^2\leqslant\chi^2_\alpha(2n)$是等价的.

同理可得其余两种检验问题的检验方法，结果列于表 4.7.

表 4.7　指数分布参数 θ 的假设检验法(显著性水平 α，检验统计量 $\chi^2=2n\bar{X}/\theta_0$)

原假设	备择假设	拒绝域
$\theta\leqslant\theta_0$	$\theta>\theta_0$	$\chi^2\geqslant\chi^2_{1-\alpha}(2n)$
$\theta\geqslant\theta_0$	$\theta<\theta_0$	$\chi^2\leqslant\chi^2_\alpha(2n)$
$\theta=\theta_0$	$\theta\neq\theta_0$	$\chi^2\leqslant\chi^2_{\alpha/2}(2n)$或 $\chi^2\geqslant\chi^2_{1-\alpha/2}(2n)$

继续讨论例 4.10. 假定 $\theta_0=3000$h，若加速寿命试验中 20 件受检样品的平均失效时间为 237h，问在 0.1 显著性水平下这批电子元件能否通过检验? 于是，检验原假设为 $H_0:\theta\geqslant\theta_0=3000$. 由上述讨论，检验的 MATLAB 数据处理如下.

```
clear
theta0 = 3000; % 待检验参数值
alpha = 0.1; % 显著性水平
n = 20; % 样本容量
EoLife = 237; % 加速寿命试验中样品平均失效时间
```

```
x2stat = 2 * n * (10 * EoLife)/theta0 % 检验统计量的观测值
c = chi2inv(alpha, 2 * n) % 求拒绝域临界值
if x2stat<=c % 检验决策, h=1(0)拒绝(接受)原假设
h = 1
else
h = 0
end
```

上述指令的运行结果是：

```
x2stat =
       31.6000
c =
       29.0505
h =
       0
```

结果表明，$\chi^2 > \chi_\alpha^2(2n)$，$h=0$，不能拒绝原假设，即这批电子元件应当通过检验.

4.2.3.2 非正态变量均值的大样本检验方法

前面介绍了两点分布参数 p、泊松分布参数 λ 和指数分布参数 θ 的小样本检验方法，细心的读者可能发现，这三种非正态分布有一个共同的特点，就是它们的数学期望等于分布参数. 因此，所谓分布参数的检验实质上是变量均值的检验.

对于非正态变量均值的检验，更一般的做法是进行大样本近似检验. 其一般描述是：设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$，记 $E(X)=\mu$，μ_0 是 μ 的先验取值，检验问题有如下三类.

① $H_0: \mu \leqslant \mu_0$；$H_1: \mu > \mu_0$.

② $H_0: \mu \geqslant \mu_0$；$H_1: \mu < \mu_0$.

③ $H_0: \mu = \mu_0$；$H_1: \mu \neq \mu_0$.

检验统计量取 $\bar{X}$. 由概率极限定理可知，当样本容量 n 很大时，$\bar{X}$ 近似服从 $N(\mu, S^2/n)$，其中 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 是样本方差. 实际检验中，在 $\mu=\mu_0$ 的假定下使用统计量

$$U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \stackrel{\cdot}{\sim} N(0,1),$$

于是，三类检验问题的拒绝域分别是

① $W = \{U \geqslant u_{1-\alpha}\}$；

② $W=\{U\leqslant u_{\alpha}\}$；

③ $W=\{|U|\geqslant u_{1-\alpha/2}\}$.

如果 $\mathrm{Var}(X)=\sigma^2$ 已知，则检验时可用 σ 替代S，检验统计量的分布与拒绝域的形式不变.

【例4.11】 从某一试验物中随机地抽取50个样品，测得样品的发热量(单位：J)数据记录如下：

11786, 12032, 11666, 12118, 11955, 12282, 12277, 11728, 12244, 11645,
12112, 12116, 11680, 12158, 11932, 11773, 12117, 12014, 12153, 11689,
11882, 11767, 12059, 11716, 11968, 11704, 11654, 11668, 11755, 11969,
12060, 11969, 12028, 11856, 12110, 11712, 11976, 12288, 11841, 11967,
12173, 11831, 12100, 12205, 12066, 12201, 12243, 12251, 12072, 12027.

试问，以0.05的显著性水平是否可以认为发热量的期望值是12000?

分析 依题意，检验假设是 $H_0:\mu=\mu_0$；$H_1:\mu\neq\mu_0$. 检验统计量 $U=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\dot{\sim}N(0,1)$，拒绝域的形式为 $|U|\geqslant u_{1-\alpha/2}$.

MATLAB数据处理

```
clear
load frl % 预先编写数据文件 frl.mat，并存放到当前工作路径下
alpha = 0.05; % 显著性水平
mu0 = 12000; % 待检验参数值
U = (mean(frl) - mu0)/(std(frl)/sqrt(length(frl))) % 检验统计量的观测值
c = norminv(1 - alpha/2,0,1) % 求拒绝域临界值
if abs(U)> = c % 检验决策，h=1(0)拒绝(接受)原假设
h = 1
else
h = 0
end
```

上述指令的运行结果是：

```
U =
    -0.9985
c =
      1.9600
h =
      0
```

结果表明不能拒绝原假设，试验物的发热量符合期望值.

实际上，由于大样本均值检验为 U 检验，故可直接调用 MATLAB 的 U 检验函数 ztest，需要注意的是要用样本标准差 std(x)代替正态分布的标准差 sigma 作为输入参数.

```
[h,p,Mci,U] = ztest(fr1,mu0,std(fr1),alpha)
```

上述指令的运行结果是：

```
h =

     0

p =

     0.3180

Mci =

     11917       12027

U =

    -0.9985
```

显然，计算出的统计量 U 的观测值与检验结论同前面一致.

在大样本均值检验问题中，一个重要的应用是两个比率的比较. 其一般描述如下.

设 $X_1,X_2,\cdots,X_n$ i.i.d $\sim b(1,p_1)$，$Y_1,Y_2,\cdots,Y_m$ i.i.d. $\sim b(1,p_2)$，两样本独立，需对 p_1 与 p_2 进行比较，这等价于下列三种假设检验问题之一.

① $H_0:p_1\leqslant p_2;H_1:p_1>p_2$.

② $H_0:p_1\geqslant p_2;H_1:p_1<p_2$.

③ $H_0:p_1=p_2;H_1:p_1\neq p_2$.

由概率极限定理可知，当样本容量 n 很大时，在 $p_1=p_2$ 的假定下，检验统计量

$$U=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\hat{p}(1-\hat{p})}}\dot{\sim}N(0,1),$$

其中 $$\hat{p}_1=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \hat{p}_2=\frac{1}{m}\sum_{i=1}^{m}Y_i,\quad \hat{p}=\frac{n\hat{p}_1+m\hat{p}_2}{n+m}.$$

于是，三类检验问题的拒绝域分别是

① $W=\{U\geqslant u_{1-\alpha}\}$;

② $W=\{U\leqslant u_{\alpha}\}$;

③ $W=\{|U|\geqslant u_{1-\alpha/2}\}$.

【例 4.12】 女性色盲的比例比男性低，从随机抽取的 467 名男性中发现有 8 名色盲，而 433 名女性中发现 1 人色盲，在 0.01 显著性水平下能否认为女性色盲的比例比男性低？

分析 设男性色盲的比例为 p_1，女性色盲的比例为 p_2，那么要检验的假设为 H_0：$p_1 \geqslant p_2$；H_1：$p_1 < p_2$.

MATLAB 数据处理

```
clear
alpha = 0.01; % 显著性水平
ESTp1 = 8/467;
ESTp2 = 1/433;
ESTp = (8 + 1)/(467 + 433);
U = (ESTp1 - ESTp2)/sqrt((1/467 + 1/433) * ESTp * (1 - ESTp)) % 检验统计量的观测值
c = norminv(alpha, 0, 1) % 求拒绝域临界值
if U< = c % 检验决策, h=1(0)拒绝(接受)原假设
h = 1
else
h = 0
end
```

上述指令的运行结果是：

```
U =
    2.2328
c =
   -2.3263
h =
    0
```

结果表明，在 0.01 显著性水平下不能拒绝原假设，即可以认为女性色盲的比例比男性低.

4.3 变量分布形态的检验

通过前几节的讨论，我们已经了解了假设检验的基本思想，并讨论了当分布形式已知时关于其中未知参数的假设检验问题. 然而，可能遇到这样的情形，如例 4.6 中，认定标准方法下的钢的产率服从正态分布通常是合理的，但是新操作方法下钢的产率是否仍服从正态分布是需要斟酌的，因为影响钢的产率的条件毕竟发生了改变. 因此在例 4.6 问题的分析中，更为严谨的思考应当包括识别新操作方法下钢的产率是否为某个正

态变量. 此类问题通常称为变量分布形态的检验, 属于非参数检验问题. 本节讨论非参数检验的几个基本方法及其应用.

4.3.1 K.Pearson-Fisher 检验

K.Pearson-Fisher 检验是非参数检验的基本方法, 主要有两个方面的应用: 一是关于变量分布形态拟合优度检验, 通常称为 χ^2 拟合优度检验; 另一是关于二维变量独立性的检验, 通常称为列联表的独立性检验.

4.3.1.1 χ^2 拟合优度检验

χ^2 拟合优度检验是关于变量 X 分布形态的某种先验知识或猜测是否为真的统计推断方法. 记变量 X 的分布函数为 $F_X(x)$, $F_0(x;\theta_1,\theta_2,\cdots,\theta_r)$ 是关于 $F_X(x)$ 的先验知识或猜测, 则 χ^2 拟合优度检验的假设是

$$H_0: F_X(x)=F_0(x;\theta_1,\theta_2,\cdots,\theta_r); \quad H_1: F_X(x)\neq F_0(x;\theta_1,\theta_2,\cdots,\theta_r).$$

在应用中, 若 X 为离散型变量, 则 H_0 可转述为概率函数的表达; 若 X 为连续型变量, 则 H_0 可转述为概率密度函数的表达.

在对上述假设 H_0 进行 χ^2 检验时, 总是假定 $F_0(x;\theta_1,\theta_2,\cdots,\theta_r)$ 的理论形式是已知的, 但其参数未知. 因此, 应用中 χ^2 拟合优度检验法包括两个环节: 先用极大似然估计法估计分布参数 $\hat{\theta}_{\mathrm{MLE}}=(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_r)$, 然后再对假设 $H_0: F_X(x)=F_0(x;\theta_1,\theta_2,\cdots,\theta_r)$ 进行检验.

仅对 χ^2 拟合优度检验法的步骤说明如下.

① 分割 X 的取值范围. 将变量 X 的取值范围分成 k 个互不重叠的小区间, 记作

$$A_1=[a_0,a_1),\ A_2=[a_1,a_2),\ \cdots,\ A_k=[a_{k-1},a_k),$$

这些区间的长度可以不等.

② 统计样本数据 $(x_1,\ x_2,\ \cdots,\ x_n)$ 落入第 i 个小区间 A_i 的实测频数 f_i. 注意: $\sum_{i=1}^{k} f_i = n$.

③ 计算变量 X 落入第 i 个小区间 A_i 的理论频数 $n\hat{p}_i$. 其中, 变量 X 落入第 i 个小区间 A_i 的概率

$$\hat{p}_i=F_0(a_i;\ \hat{\theta}_1,\ \hat{\theta}_2,\ \cdots,\ \hat{\theta}_r)-F_0(a_{i-1};\ \hat{\theta}_1,\ \hat{\theta}_2,\ \cdots,\ \hat{\theta}_r) \quad (i=1,2,\cdots,k).$$

注意到分布参数是由极大似然法估计出的, 因此这个概率本质上也是一个估计值.

④ 计算检验统计量 $\chi^2=\sum_{i=1}^{k}\frac{(f_i-n\hat{p}_i)^2}{n\hat{p}_i}$ 的值. 这个统计量最初是由 K.Pearson 在

1900年引进的，K.Pearson证明了

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}\overset{L}{\sim}\chi^2(k-1),$$

其中 p_i 为依据分布函数$F_0(x;\theta_1,\theta_2,\cdots,\theta_r)$计算的变量 X 落入第i 个小区间A_i 的理论概率. 这一结论后由 Fisher 改进，他证明了若分布函数 $F_0(x;\theta_1,\theta_2,\cdots,\theta_r)$中的参数 $\theta_1,\theta_2,\cdots,\theta_r$ 由其极大似然估计代替得$F_0(x;\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_r)$，则当样本容量 $n\to\infty$ 时

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-n\hat{p}_i)^2}{n\hat{p}_i}\overset{L}{\sim}\chi^2(k-r-1),$$

证明参见文献[7]. 因此，实际应用中一般要求 $n\geqslant 50$，以及每一个 $n\hat{p}_i$ 都不小于5. 否则应适当合并区间，使 $n\hat{p}_i$ 满足这个要求.

⑤ 作出检验决策. 显然，在 χ^2 统计量中各个实测频数 f_i 与理论频数$n\hat{p}_i$ 之间偏差平方的大小标志着经验分布与理论分布之间差异的大小. 如果 χ^2 统计量的值过于偏大，则表明样本信息不支持原假设 H_0 成立的假定，因此对于给定的显著性水平 α，检验准则为 $P\{\chi^2>\chi^2_{1-\alpha}(k-r-1)\}\leqslant\alpha$，即当检验统计量的实测值 $\chi^2>\chi^2_{1-\alpha}(k-r-1)$时，则在显著性水平 α 下拒绝原假设H_0，否则保留 H_0.

下面举例说明 χ^2 拟合优度检验法的应用.

【例4.13】 表4.8中数据是200个零件的直径 X(单位：cm).

表4.8

直径	2.25	2.35	2.45	2.55	2.65	2.75	2.85	2.95
频数	3	4	5	11	12	17	19	26
直径	3.05	3.15	3.25	3.35	3.45	3.55	3.65	3.75
频数	24	22	19	13	13	7	3	2

能否验证直径 X 服从正态分布?

分析 依题意检验的假设是 H_0: 零件直径 X 服从正态分布$N(\mu,\sigma^2)$. 其中，参数 μ,σ^2 均未知. 因此，首先要求出参数 μ,σ^2 的极大似然估计：

$$\hat{\mu}_{\mathrm{MLE}}=\frac{1}{n}\sum_{i=1}^{k}f_ix_i \quad (\text{分组数据的样本均值});$$

$$\hat{\sigma}^2_{\mathrm{MLE}}=\frac{1}{n}\sum_{i=1}^{k}f_ix_i^2-\mu^2_{\mathrm{MLE}} \quad (\text{分组数据的样本方差}).$$

然后在按照前述①～⑤步进行 χ^2 拟合优度检验.

MATLAB 数据处理

① 输入原始数据，并求分布参数的极大似然估计.

```
clear
x = [2.25, 2.35, 2.45, 2.55, 2.65, 2.75, 2.85, 2.95, 3.05, 3.15, 3.25, 3.35, 3.45, 3.55, 3.65, 3.75];
f = [3, 4, 5, 11, 12, 17, 19, 26, 24, 22, 19, 13, 13, 7, 3, 2];
n = sum(f);
MU = sum(f. * x)./n
SIGMA = sqrt(sum(f. * (x.^2))./n - MU.^2)
```

上述指令的运行结果是：

```
MU =
    3.0090
SIGMA =
    0.3210
```

根据计算结果，检验的原假设修正为 $H_0: X \sim N(3.009, 0.3210^2)$.

② 样本数据分组.

题目给出的数据已是分组数据，共分为 16 组，且每组的频数已经统计出．但是，前 3 组数据和后 3 组数据的频数偏小，故分别将前、后 3 组数据进行合并，这样可得 12 组数据．这 12 组数据所属的数据组的区间边界值如下.

```
a = [];
for k = 1: 11
aa = (x(2 + k) + x(3 + k))/2; % 小区间边界点取相邻两个数据的中点
a = [a, aa];
end
a = [ - inf, a, inf]' % 由于正态变量在整个数轴上取值，最小边界点为 - ∞，最大边界点为 + ∞
```

上述指令的运行结果是：

```
a =
    - Inf
   2.5000
   2.6000
   2.7000
   2.8000
   2.9000
```

```
    3.0000
    3.1000
    3.2000
    3.3000
    3.4000
    3.5000
       Inf
```

③ 统计经验频数.

经验频数题目已经给出，只需分别合并前、后 3 组的频数.

```
f = [f(1) + f(2) + f(3),f(4:13),f(14) + f(15) + f(16)]'
```

上述指令的运行结果是：

```
f =
    12
    11
    12
    17
    19
    26
    24
    22
    19
    13
    13
    12
```

④ 计算理论频数.

```
PEST = [];
for i = 1:12
pp = normcdf(a(i + 1),MU,SIGMA) - normcdf(a(i),MU,SIGMA);
PEST = [PEST,pp];
end
THEF = n * PEST'
```

上述指令的运行结果是：

```
THEF =
   11.2776
```

```
     8.9789
    13.3124
    17.9255
    21.9214
    24.3472
    24.5591
    22.4988
    18.7193
    14.1449
     9.7072
    12.6077
```

⑤ 计算检验统计量的观测值.

```
CHI2EST = sum((f - THEF).^2./THEF)
```

上述指令的运行结果是:

```
CHI2EST =
     2.4469
```

⑥ 检验决策.

```
k = 12;
r = 2;
alpha = 0.05;
df = k - r - 1;
REFCR = chi2inv(1 - alpha, df); % 拒绝域临界值
p = 1 - chi2cdf(CHI2EST, df); % 检验的 p 值
if CHI2EST>REFCR
h = 1;
else
h = 0;
end
alpha, h, p
stat = [k, r, CHI2EST, REFCR]
```

上述指令的运行结果是:

```
alpha =
     0.0500
h =
```

```
        0
p =
        0.9823
stat =
        12.0000      2.0000      2.4469      16.9190
```

计算结果表明，在 0.05 显著性水平下，h=0 保留原假设 H_0，即 χ^2 拟合优度检验认为零件直径 $X \sim N(3.009, 0.1030)$．最小显著性概率 p=0.9823 表明，当前样本数据下不能拒绝原假设 H_0 的置信程度高达 98%．

【例 4.14】 在 20 天内，从维尼纶正常生产时的生产报表中看到的维尼纶纤度(纤维的粗细程度的一种度量)的情况，有如下 100 个数据：

1.36, 1.49, 1.43, 1.41, 1.37, 1.40, 1.32, 1.42, 1.47, 1.39,
1.41, 1.36, 1.40, 1.34, 1.42, 1.42, 1.45, 1.35, 1.42, 1.39,
1.44, 1.42, 1.39, 1.42, 1.42, 1.30, 1.34, 1.42, 1.37, 1.36,
1.37, 1.34, 1.37, 1.37, 1.44, 1.45, 1.32, 1.48, 1.40, 1.45,
1.39, 1.46, 1.39, 1.53, 1.36, 1.48, 1.40, 1.39, 1.38, 1.40,
1.36, 1.45, 1.50, 1.43, 1.38, 1.43, 1.41, 1.48, 1.39, 1.45,
1.37, 1.37, 1.39, 1.45, 1.31, 1.41, 1.44, 1.44, 1.42, 1.47,
1.35, 1.36, 1.39, 1.40, 1.38, 1.35, 1.42, 1.43, 1.42, 1.42,
1.42, 1.40, 1.41, 1.37, 1.46, 1.36, 1.37, 1.27, 1.37, 1.38,
1.42, 1.34, 1.43, 1.42, 1.41, 1.41, 1.44, 1.48, 1.55, 1.37.

正常情况下，维尼纶纤度服从正态分布．试根据这 100 个样本数据在 0.10 显著性水平下验证生产是正常的．

分析 这是一个正态拟合检验问题．检验的原假设是 H_0：维尼纶纤度 X 服从正态分布 $N(\mu, \sigma^2)$．其中，参数 μ, σ^2 均未知．

MATLAB 数据处理

① 输入原始数据，进行未知参数的极大似然估计．

```
clear
load wnlxd; % 预先编写数据文件 wnlxd.mat，并存放到当前工作路径下
n = length(wnlxd);
[MU, SIGMA] = normfit(wnlxd)
```

上述指令的运行结果是：

```
MU =
   1.4042
```

```
SIGMA =
   0.0478
```

于是，检验假设修正为 H_0：$X \sim N(1.4042, 0.0178^2)$.

② 样本数据分组.

```
[f, med] = hist(wnlxd);
F_MED = [f', med']
```

上述指令的运行结果是：

```
F_MED =
      1     1.2840
      4     1.3120
      7     1.3400
     22     1.3680
     23     1.3960
     20     1.4240
     13     1.4520
      7     1.4800
      1     1.5080
      2     1.5360
```

利用 hist 指令自动分为 10 分组，并统计各组频数．由计算结果可知，前 3 组数据和后 3 组数据的频数偏小，故分别将前、后 3 组数据进行合并，这样可得 6 组数据．这 6 组数据所属的数据组的区间边界值如下．

```
a = [];
for k = 1:5
aa = (med(2 + k) + med(3 + k))/2;
a = [a,aa];
end
a = [ - inf,a,inf]'
```

上述指令的运行结果是：

```
a =
     - Inf
    1.3540
    1.3820
    1.4100
    1.4380
```

```
   1.4660
      Inf
```

③ 统计经验频数.

经验频数②已经给出，只需分别合并前、后 3 组的频数.

```
f = [f(1) + f(2) + f(3),f(4:7),f(8) + f(9) + f(10)]'
```

上述指令的运行结果是：

```
f =
    12
    22
    23
    20
    13
    10
```

④ 计算理论频数.

```
PEST = [];
for i = 1:6
pp = normcdf(a(i + 1),MU,SIGMA) - normcdf(a(i),MU,SIGMA);
PEST = [PEST,pp];
end
THEF = n * PEST'
```

上述指令的运行结果是：

```
THEF =
   14.6617
   17.4417
   22.7293
   21.2101
   14.1726
    9.7845
```

⑤ 计算检验统计量的观测值.

```
CHI2EST = sum((f - THEF).^2./THEF)
```

上述指令的运行结果是：

```
CHI2EST =
    1.8485
```

⑤ 检验决策.

```
k = 6;
r = 2;
alpha = 0.10;
df = k - r - 1;
REFCR = chi2inv(1 - alpha, df); % 拒绝域临界值
p = 1 - chi2cdf(CHI2EST, df); % 检验的 p 值
if CHI2EST > REFCR
h = 1;
else
h = 0;
end
alpha, h, p, CHI2EST, REFCR
```

上述指令的运行结果是：

```
alpha =
     0.1000
h =
     0
p =
     0.6044
CHI2EST =
     1.8485
REFCR =
     6.2514
```

计算结果表明，在 0.10 显著性水平下，h=0 保留原假设 H_0，即 χ^2 拟合优度检验认为维尼纶纤度 $X \sim N(1.4042, 0.0178^2)$. 由最小显著性概率 p=0.6044 表明，当前样本数据下不能拒绝原假设，H_0 有较高的可信程度.

4.3.1.2 列联表的独立性检验

K.Pearson-Fisher 的 χ^2 统计量有一个很特别的应用，即可以用来检验两个分类变量的独立性.

设 X 与 Y 是两个分类变量，不妨设 X 有 s 个类别 $A_1, A_2, \cdots, A_s$，Y 有 t 个类别 $B_1, B_2, \cdots, B_t$，将被调查的 n 个样品按其所属类别进行分类，列成如下一张 $s \times t$ 的二维表，见表 4.9.

表 4.9 也称为 $s \times t$ 列联表. 其中，f_{ij} 表示同时具有属性 A_i 和 B_j 的样品频数($i=1$,

$2,\cdots,s;\ j=1,2,\cdots,t$)，$f_{i\cdot}=\sum_{j=1}^{t}f_{ij}$，$f_{\cdot j}=\sum_{i=1}^{s}f_{ij}$，$\sum_{i=1}^{s}\sum_{j=1}^{t}f_{ij}=\sum_{i=1}^{s}f_{i\cdot}=\sum_{j=1}^{t}f_{\cdot j}=n$.

表 4.9　　$s\times t$ 列联表

	B_1	B_2	$\cdots$	B_t	Σ
A_1	f_{11}	f_{12}	$\cdots$	f_{1t}	$f_{1\cdot}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2t}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_s	f_{s1}	f_{s2}	$\cdots$	f_{st}	$f_{s\cdot}$
Σ	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot t}$	n

用 K.Pearson-Fisher 的 χ^2 统计量来检验变量 X 与 Y 的独立性，检验假设是 H_0:X 与 Y 是独立的；H_1:X 与 Y 不独立.

对 H_0 的检验依赖表 4.9，因此这类问题亦称为列联表的独立性检验. 记

$$p_{ij}=P\{X\in A_i,\ Y\in B_j\},\quad p_{i\cdot}=\sum_{j=1}^{t}p_{ij}=p\{X\in A_i\},\quad p_{\cdot j}=\sum_{i=1}^{s}p_{ij}=P\{Y\in B_j\},$$

其中 $i=1,2,\cdots,s$，$j=1,2,\cdots,t$. 于是检验假设可进一步明确为

H_0:$p_{ij}=p_{i\cdot}p_{\cdot j}$，对所有 i，j 均成立；

H_1:$p_{ij}\neq p_{i\cdot}p_{\cdot j}$，至少存在一对 i，j 使之成立.

又记 p_{ij}，$p_{i\cdot}$和 $p_{\cdot j}$的极大似然估计分别为$\hat{p}_{ij}$，$\hat{p}_{i\cdot}$和$\hat{p}_{\cdot j}$，并且

$$\hat{p}_{ij}=f_{ij}/n,\quad \hat{p}_{i\cdot}=f_{i\cdot}/n,\quad \hat{p}_{\cdot j}=f_{\cdot j}/n.$$

因此，对 H_0 的检验可以通过分析偏差平方和 $\sum_{i=1}^{s}\sum_{j=1}^{t}(\hat{p}_{ij}-\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2$ 得到，当 H_0 成立时这个偏差平方和不应过分偏大. 基于这种理解，可得 K.Pearson-Fisher 的 χ^2 统计量的变式表达为

$$\chi^2=\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(f_{ij}-n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}\sim\chi^2((s-1)(t-1)),$$

当 H_0 成立时 χ^2 统计量的观测值不应过分偏大. 于是，对于给定的显著性水平 α，检验准则为

$$P\{\chi^2>\chi^2_{1-\alpha}((s-1)(t-1))\}\leqslant\alpha,$$

即当检验统计量的实测值 $\chi^2>\chi^2_{1-\alpha}((s-1)(t-1))$时，则在显著性水平 α 下拒绝原假设H_0. 否则保留 H_0. 在 χ^2 统计量观测值的计算中注意，$n\hat{p}_{i\cdot}\hat{p}_{\cdot j}=f_{i\cdot}f_{\cdot j}/n$ ($i=1,2,\cdots,s;\ j=1,2,\cdots,t$).

下面举例说明列联表的独立性检验.

【例 4.15】 某地调查了 3000 名失业人员，按性别和文化程度分类如表 4.10 所示.

表 4.10

	大专以上	中专技校	高中	初中及以下	合计
男	40	138	620	1043	1841
女	20	72	442	625	1159
合计	60	210	1062	1668	3000

试在 0.05 显著性水平下检验失业人员的性别与文化程度是否有关.

分析 这是列联表的独立性检验问题. 检验原假设为 H_0:失业人员的性别与文化程度无关.

MATLAB 数据处理

```
clear
alpha = 0.05;
f = [40,138,620,1043;20,72,442,625];
[s,t] = size(f); % 提取列联表的行、列数
df = (s - 1) * (t - 1);
fi_ = sum(f'); % 行边际频数
f_j = sum(f); % 列边际频数
n = sum(sum(f));
nfi_f_j = zeros(s,t);
for i = 1:2
for j = 1:4
nfi_f_j(i,j) = fi_(i) * f_j(j)/n; % 联合分布律
end
end
CHI2EST = sum(sum((f - nfi_f_j).^2./nfi_f_j)); % 检验统计量的值
REFCR = chi2inv(1 - alpha,df); % 拒绝域临界值
p = 1 - chi2cdf(CHI2EST,df); % 检验的 p 值
if CHI2EST>REFCR
h = 1;
else
h = 0;
end
```

```
alpha,h,p,CHI2EST,REFCR
```

上述指令的运行结果是：

```
alpha =
      0.0500
h =
      0
p =
      0.0620
CHI2EST =
      7.3320
REFCH =
      7.8147
```

计算结果表明，在 0.05 显著性水平下，h=0，p>alpha 不能拒绝原假设，即认为失业人员的性别与文化程度无关.

4.3.2　Колмогоров-Смирнов 检验

假设变量 X 的分布函数 $F(x)$ 连续但未知，在给定显著性水平 α 下，要检验假设

$$H_0: F(x) = F_0(x);\quad H_1: F(x) \neq F_0(x).$$

这个问题可以用 χ^2 拟合优度检验法来检验.

但是，χ^2 拟合优度检验的实质是比较样本频率 $\frac{\nu_i}{n}$ 与理论频率 $\hat{p}_i = F_0(a_i) - F_0(a_{i-1})$，也就是说只是检验了

$$H_0: F(a_i) - F(a_{i-1}) = F_0(a_i) - F_0(a_{i-1})\quad (i = 1, 2, \cdots, k),$$

其中 a_i 是在连续变量离散化的区间划分过程中得到的，也就是说只是检验了在区间的分点处 H_0 是否成立而已，这样导致了纳伪风险的增加. 于是，人们转而研究更加完善的检验方法.

早在 20 世纪 30 年代初，Колмогоров 对分布拟合优度检验问题进行了深人的研究，得到了 Колмогоров 定理，进而建立了分布拟合优度检验问题的 Колмогоров 检验法和 Смирнов 检验法.

4.3.2.1　Колмогоров 检验法

Колмогоров 检验法也是比较样本经验函数 $F_n(x)$ 和变量分布函数 $F_0(x)$ 的. 但它不是在划分的区间上考虑 $F_n(x)$ 与原假设的分布函数 $F_0(x)$ 之间的偏差，而是在每一点上考虑它们之间的偏差. 这就克服了 χ^2 检验法依赖于区间划分的缺点，但其应用范

围要窄一些，仅适应于变量的分布函数是连续函数的情形.

根据 Гливенко 定理，当 n 充分大时，样本经验分布函数 $F_n(x)$是变量的分布函数 $F(x)$的很好近似，$F_n(x)$与 $F(x)$的偏差一般不应太大. Колмогоров 用 $F_n(x)$与 $F(x)$之间的偏差的最大值构造一个统计量

$$D_n = \sup_{-\infty<x<+\infty} \left| F_n(x) - F_0(x) \right|,$$

并且得到了下面的定理.

定理 4.1（Колмогоров 定理）　设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim F(x)$ $(n = 1, 2, \cdots)$，$F(x)$为连续的分布函数，在 $F(x) = F_0(x)$(已知)的条件下，有

$$\lim_{n\to\infty} P\left\{ D_n < \frac{x}{\sqrt{n}} \right\} = K(x),$$

其中

$$K(x) = \begin{cases} \sum\limits_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2}, & x > 0, \\ 0, & x \leqslant 0 \end{cases}$$

称为 Колмогоров 分布.

定理的证明参见文献[7].

根据定理 4.1 检验 $H_0: F(x) = F_0(x)$，若假定 H_0 为真，则当 n 充分大时，检验统计量 $D_n = \sup\limits_{-\infty<x<+\infty} \left| F_n(x) - F_0(x) \right|$的值一般应该比较小，若 D_n 的值较大就应该拒绝 H_0. 于是，对给定的显著性水平 α，拒绝域形式为 $D_n \geqslant c$，检验准则为求满足条件 $P(D_n \geqslant c \mid H_0 \text{为真}) \leqslant \alpha$ 的拒绝域临界值 c.

记 $D_{n,1-\alpha}$ 为 Колмогоров 分布的上侧 α 分位数，即 $P\{D_n \geqslant D_{n,1-\alpha}\} = \alpha$，则 Колмогоров 检验法的决策法则是：根据样本数据计算出检验统计量 D_n 的观测值，若

① 当 $D_n \geqslant D_{n,1-\alpha}$时，拒绝 H_0，即认为 $F(x) \neq F_0(x)$；

② 当 $D_n < D_{n,1-\alpha}$时，接受 H_0，即认为 $F(x) = F_0(x)$.

应用 Колмогоров 检验法时，原假设 $H_0: F(x) = F_0(x)$中的 $F_0(x)$的参数应该是已知的. 当参数未知时，对于正态分布或指数分布，可用参数的大样本估计代替，不过此时的检验是近似的，且显著性水平 α 在 0.1～0.2 为宜.

下面概括地给出在显著性水平 α 下，用 Колмогоров 检验法检验假设

$$H_0: F(x) = F_0(x); \quad H_1: F(x) \neq F_0(x)$$

的步骤，其中分布函数 $F(x)$是连续函数.

① 样本数据排序. 将样本数据 $x_1, x_2, \cdots, x_n$(通常 $n \geqslant 50$)按由小到大的次序排列得 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$.

② 求出经验分布函数.

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{1}{n}\sum_{i=1}^{k}\nu_i, & x_{(k)}\leqslant x<x_{(k+1)}(k=1,2,\cdots,n-1),\\ 1, & x\geqslant x_{(n)},\end{cases}$$

其中 ν_i 为样本数据 $x\in[x_{(i)},x_{(i+1)})$ 的频数，且 $\sum\nu_i=n$.

③ 计算检验统计量 D_n 的值.

$$D_n=\sup_{-\infty<x<+\infty}|F_n(x)-F_0(x)|=\max_{\forall i}\{|F_n(x_{(i)})-F_0(x_{(i)})|,|F_n(x_{(i+1)})-F_0(x_{(i)})|\},$$

其中，规定 $F_n(x_{(n+1)})=1$.

④ 求 Колмогоров 分布的上侧 α 分位数 $D_{n,1-\alpha}$. 当 $n>100$ 时，常用的 $D_{n,1-\alpha}$ 近似公式如下：

$$D_{n,0.80}\approx1.07/\sqrt{n},\quad D_{n,0.90}\approx1.23/\sqrt{n},\quad D_{n,0.95}\approx1.36/\sqrt{n},\quad D_{n,0.99}\approx1.63/\sqrt{n}.$$

⑤ 检验决策.

若 $D_n\geqslant D_{n,1-\alpha}$，则拒绝 H_0，认为样本数据非来自理论分布 $F_0(x)$ 的；

若 $D_n<D_{n,1-\alpha}$，则接受 H_0，认为样本数据是来自理论分布 $F_0(x)$ 的.

4.3.2.2 Смирнов 检验法

Смирнов 检验法是对 Колмогоров 检验法的一种推广.

设 $X_1,X_2,\cdots,X_n$ i.i.d $\sim F(x)$, $Y_1,Y_2,\cdots,Y_m$ i.i.d. $\sim G(x)$ $(n,m=1,2,\cdots)$, $F(x)$ 和 $G(x)$ 均为连续的分布函数，$-\infty<x<+\infty$，在显著性水平 α 下，检验假设

$$H_0:F(x)=G(x);\quad H_1:F(x)\neq G(x).$$

用 $F_n(x)$ 和 $G_m(x)$ 分别表示两样本的经验分布函数，用它们构造检验统计量

$$D_{nm}=\sup_{-\infty<x<+\infty}|F_n(x)-G_m(x)|,$$

Смирнов 证明了下面的定理.

定理 4.2（Колмогоров-Смирнов 定理） 当 H_0 为真且样本容量 n 和 m 分别趋向于 ∞ 时，有

$$\lim_{n,m\to\infty}P\left\{\sqrt{\frac{nm}{n+m}}D_{nm}<x\right\}=K(x),$$

其中 $K(x)$ 是 Колмогоров 分布函数.

根据定理 4.2，可得检验 $H_0:F(x)=G(x)$ 的 Смирнов 检验法则(近似)：

① 若 $D_{nm}\geqslant D_{nm,1-\alpha}$，则拒绝 H_0，认为 $F(x)\neq G(x)$；

② 若 $D_{nm}<D_{nm,1-\alpha}$，则接受 H_0，认为 $F(x)=G(x)$.

应用中，确定 Колмогоров 分布的分位数 $D_{nm,1-\alpha}$ 时，用 $N=\left[\frac{nm}{n+m}\right]$ 代替前述分位

数近似公式中的 n，而计算 D_{nm} 的观测值用公式

$$D_{nm}=\max_{\forall i}\left|F_n(x_{(i)})-G_m(x_{(i)})\right|,$$

其中，x_i 为划分变量值域的第 i 个小区间的组中值.

MATLAB 将这两种检验方法统称为 Колмогоров-Смирнов(英文书写为 Kolmogorov-Smirnov)检验，并提供了两个检验函数 kstest 和 kstest2.

① kstest.

函数 kstest 用于大样本情形下连续变量分布形态的拟合优度检验. 调用格式为

```
[h,p,stats,cv] = kstest(x,cdf,alpha,tail)
```

其中，输入参数 x 为样本数据向量，cdf 为检验的原假设所指定的分布形式(具体引用为变量的累积分布函数，缺省时 cdf=[]，表示拟合标准正态分布)，alpha 为检验的显著性水平(缺省时为 0.05)，tail 为备择假设类型的标示值. 输出参数 h 为检验决策，p 为拒绝原假设的最小显著性概率，stats 为检验统计量的值，cv 为拒绝域的临界值.

② 函数 kstest2.

函数 kstest2 用于大样本情形下两个连续变量分布一致性的检验. 调用格式为

```
[h,p,stats] = kstest2(x,y,alpha,tail)
```

检验的原假设是两个变量服从相同的分布. 输入参数 x 和 y 分别为两个样本的数据向量，其他输入、输出参数的意义同 kstest.

【例 4.16】 在 0.10 显著性水平下，用 Колмогоров-Смирнов 检验法对例 4.14 中的维尼纶纤度数据进行正态性检验.

分析 检验的原假设是维尼纶纤度服从正态分布.

MATLAB 数据处理

```
clear
load wnlxd
[MU,SIGMA] = normfit(wnlxd)
x = (wnlxd - MU)/SIGMA;
[h,p,stats,cv] = kstest(x,[],0.10,0)
```

上述指令的运行结果是：

```
MU =
    1.4042
SIGMA =
    0.0478
h =
    0
```

```
p =
    0.3713
stats =
    0.0904
cv =
    0.1207
```

结果表明，接受原假设，即认为维尼纶纤度服从均值为 1.4042、标准差为 0.0478 的正态分布.

这里补充说明一点，关于两个变量分布一致性的检验方法，除 Смирнов 检验法，还有如 Wilcoxon 符号秩检验法、符号检验法等，有些方法在小样本条件下可能更为有效，限于篇幅本书未作介绍. 希望了解这些方法的读者可参阅其他数理统计教程，如文献[2]. 相关的 MATLAB 检验函数可参见本书附录 B.

4.3.3 正态性检验

检验变量是否服从正态分布是统计应用中最常见也是最重要的问题. 此类问题当然可以用 Колмогоров-Смирнов 检验法进行. 但是，由于受样本容量因素的影响，有时检验效果可能不理想. 因此，人们发现了一些专门的正态性检验方法，其检验效果一般比通用方法好. 这里介绍三种常用的正态性检验方法.

4.3.3.1 正态概率纸检验法

正态概率纸是一种现场统计常用的判断变量正态性的简单工具，使用它可以很快地判断变量是否服从正态分布，还能够粗略地估计出分布的数字特征.

首先介绍正态概率纸的构造原理.

设变量 X 的分布函数为 $F(x)$，需要检验

$$H_0: X \sim N(\mu, \sigma^2) \quad (-\infty < \mu < +\infty,\ \sigma^2 > 0).$$

在原假设 H_0 成立时，$\frac{X-\mu}{\sigma} = U \sim N(0,1)$，而且 $F(x)$ 可用标准正态分布 $N(0,1)$ 分布函数 $\Phi(x)$ 来表示

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(u),$$

其中

$$u = \frac{1}{\sigma}(x-\mu).$$

在 xOu 直角坐标平面上，假定横轴(x 轴)与纵轴(u 轴)的单位长度相等，函数 $u = \frac{1}{\sigma}(x-\mu)$ 的图像是一条直线，过点$(\mu, 0)$，斜率为$\frac{1}{\sigma}$.

为使这条直线能够直观地解释变量的取值 x 与 $P\{X\leqslant x\}$ 之间的关系，进行如下坐标刻度更新：在直角坐标系 xOu 中，保持横轴上 x 的刻度不变，而把纵轴上 u 的刻度更新为 $y=100\Phi(u)$，并规定 $100\Phi(-\infty)=0$，$100\Phi(+\infty)=100$. 这样就将直角坐标 xOu 更新为直角坐标系 xOy. 由于 y 轴上的刻度 0 与 100 分别对应 u 轴上的 $-\infty$ 与 $+\infty$，因此 y 轴上无法标示出 0 与 100，一般 y 轴上的刻度标示限于 0.01 到 99.99 之间. 称以直角坐标系 xOy 为刻度体系的坐标纸为**正态概率纸**.

根据正态概率纸的构造原理可知，在 xOu 直角坐标系中的 x 与 u 的关系，在 xOy 直角坐标系中就成为 x 与 $y=100P\{X\leqslant x\}(=100F(x)=100\Phi(u))$ 的关系；反之亦然. 特别对于正态概率纸上的一条直线，若该直线能表示为 $u=\dfrac{1}{\sigma}(x-\mu)$，则 $100F(x)$ 与 x 的关系为

$$100F(x)=100\Phi(u)=100\Phi\left(\frac{x-\mu}{\sigma}\right),$$

即

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right),$$

也就是说，$F(x)$ 是一个正态分布的分布函数.

这表明，正态概率纸上斜率存在且大于零的全体直线所组成的集合与全体正态分布所组成的正态分布族之间存在一一对应关系.

下面介绍正态概率纸检验法检验原假设 H_0 的具体步骤.

为了检验假设 H_0，设 $X_1,X_2,\cdots,X_n$ i.i.d. $\sim F(x)$，求出经验分布函数 $F_n(x)$，然后在正态概率纸描出点列 $(x_i,100F_n(x_i))$ $(i=1,2,\cdots,n)$. 根据 Гливенко 定理，当 n 充分大时，样本经验分布函数 $F_n(x)$ 是变量的分布函数 $F(x)$ 的很好近似. 因此，当 H_0 为真时，在正态概率纸上点列 $(x_i,100F_n(x_i))$ $(i=1,2,\cdots,n)$ 应该近似地在一条直线附近. 否则认为 H_0 不成立，即变量 X 不服从正态分布. 具体的检验步骤如下.

① 整理数据. 把样本观测值由小到大排列（设 n 个数据仅有 m 个互异），见表 4.11.

表 4.11

观测值 $x_{(i)}$	$x_{(1)}$	$x_{(2)}$	…	$x_{(m)}$
频数 ν_i	ν_1	ν_2	…	ν_m
修正经验分布函数值 $F_n^*(x_{(i)})$	$\dfrac{\nu_1}{n+1}$	$\dfrac{\nu_1+\nu_2}{n+1}$	…	$\dfrac{\nu_1+\cdots+\nu_m}{n+1}$

由于正态概率纸无法描出纵坐标为 $100F_n(x)=100$ 的点，故把 $F_n(x)$ 修正为 $F_n^*(x)$. 这种修正在样本容量比较小时很有必要；在样本容量比较大时，$F_n(x)$ 与 $F_n^*(x)$ 非常接近.

② 描点. 把点列$(x_i, 100F_n(x_i))$ $(i=1,2,\cdots,n)$描在正态概率纸上.

③ 判断. 目测这些点的位置, 如果这 m 个点近似地在一条直线 L 的附近(对应 $x_{(1)}$, $x_{(m)}$处允许偏离直线远些), 则接受原假设 H_0; 否则拒绝原假设 H_0.

④ 参数估计. 若接受原假设 H_0, 则画出这条直线 L(用最小二乘拟合, 参见第 6 章). 由 $u=\frac{1}{\sigma}(x-\mu)$可知: 当 $u=0$ 时, $x=\mu$; 当 $u=1$ 时, $\sigma=x-\mu$. 于是:

在正态概率纸上画一条水平直线 $y=50$(即 xOu 系的直线$u=0$), 它与直线 L 的交点横坐标x_0 可作为均值 μ 的估计, 即$\hat{\mu}=x_0$;

在正态概率纸上画一条水平直线 $y=84.13$(即 xOu 系的直线$u=1$), 由它与直线 L 的交点横坐标 x_1 可推出标准差 σ 的估计, 即$\hat{\sigma}=x_1-\hat{\mu}=x_1-x_0$.

在实际问题中, 通常数据都比较多, 常采用简化计算的方法: 把数据按等间隔分组, 尽量使每组至少包含一个数据, 然后以组中值作为该组所有数据的值, 每组所包含的数据个数作为取该组中值的频数, 修改的经验分布函数的观测值 $F_n^*(x_{(i)})$由组中值与它的频数决定

$$F_n^*(x_{(i)})=\frac{\nu_1+\cdots+\nu_i}{n+1}.$$

其中, n 为样本容量(数据个数); $x_{(i)}$为由小到大顺序的第 i 组的组中值; ν_i 为该组的组频数. 当数据的个数多于 50 个时, 分为 10 到 25 组为宜.

MATLAB 提供了利用正态概率纸检验变量正态性的绘图函数 normplot, 其调用格式为 normplot(x), 输入参数 x 是样本数据向量.

4.3.3.2 Lilliefors 检验

Lilliefors 检验法是对 Колмогоров 检验法的一种改进.

设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$, X 的分布未知. 需要检验

$$H_0: X\sim N(\mu,\sigma^2)\quad(-\infty<\mu<+\infty, \sigma^2>0).$$

令 $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$, $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2}$, $Z_i=\frac{X_i-\bar{X}}{S}$ $(i=1,2,\cdots,n)$, 则当 H_0 为真时, 标准化样本 $Z_1, Z_2, \cdots, Z_n$ i.i.d. $\sim N(0,1)$, 于是 Колмогоров 统计量可修正为

$$D_n=\sup_{-\infty<x<+\infty}\left|S_n(x)-\Phi(x)\right|,$$

其中, $S_n(x)$是标准化样本的经验分布函数. 这就是 Lilliefors 检验的检验统计量.

其他如检验法则、检验步骤等与 Колмогоров 检验法类似, 这里不再赘述.

由 Lilliefors 检验的检验统计量的构造特点可知, 该方法与 Колмогоров 检验法的最大不同之处是检验不需要已知分布参数, 样本的标准化避免了在正态拟合优度检验之前

对分布参数的估计，因此该方法可在小样本条件下使用.

MATLAB 提供了 Lilliefors 检验法的检验函数 lillietest，其调用格式为

```
[h, p, stats, cv] = lillietest(x, alpha, tail)
```

其输入、输出参数的意义同 kstest.

4.3.3.3 Jarque-Bera 检验

Jarque-Bera 检验是一种常用的、基于峰度与偏度联合检验的正态性检验方法.

设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim X$，X 的分布未知. 需要检验

$$H_0: X \sim N(\mu, \sigma^2) \quad (-\infty < \mu < +\infty, \sigma^2 > 0).$$

令 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$，Jarque 和 Bera 由样本峰度 $KU = \frac{B_4}{B_2^2}$ 和偏度 $SK = \frac{B_3}{B_2^{3/2}}$ 定义了如下的统计量：

$$J = \frac{n}{6}\left[SK^2 + \frac{(KU-3)^2}{4}\right],$$

并证明了在 H_0 为真的条件下，J 渐近地服从自由度为 2 的 χ^2 分布.

由于正态分布的峰度 $KU = 3$，偏度 $SK = 0$，因此检验统计量 J 的观测值越大越对 H_0 不利. 于是，对于给定的显著性水平 α，检验准则为 $P\{J > \chi^2_{1-\alpha}(2)\} \leqslant \alpha$. 当检验统计量的实测值 $J > \chi^2_{1-\alpha}(2)$ 时，则在显著性水平 α 下拒绝原假设 H_0，否则保留 H_0.

由于检验依据的是渐近分布，因此该方法应在大样本条件下使用.

MATLAB 提供了 Jarque-Bera 检验法的检验函数 jbtest，其调用格式为

```
[h, p, stats, cv] = jbtest(x, alpha, tail)
```

其输入、输出参数的意义同 kstest.

【例 4.17】 在 0.10 显著性水平下，分别用正态概率纸检验法、Lilliefors 检验法和 Jarque-Bera 检验法对例 4.14 中的维尼纶纤度数据进行正态性检验.

分析 检验的原假设是维尼纶纤度服从正态分布.

MATLAB 数据处理

```
clear
load wnlxd
```

① 正态概率纸检验法.

```
normplot(wnlxd)
```

上述指令的运行结果见图 4.1.

② Lilliefors 检验法.

```
[L_h, L_p] = lillietest(wnlxd, 0.10)
```

上述指令的运行结果是：

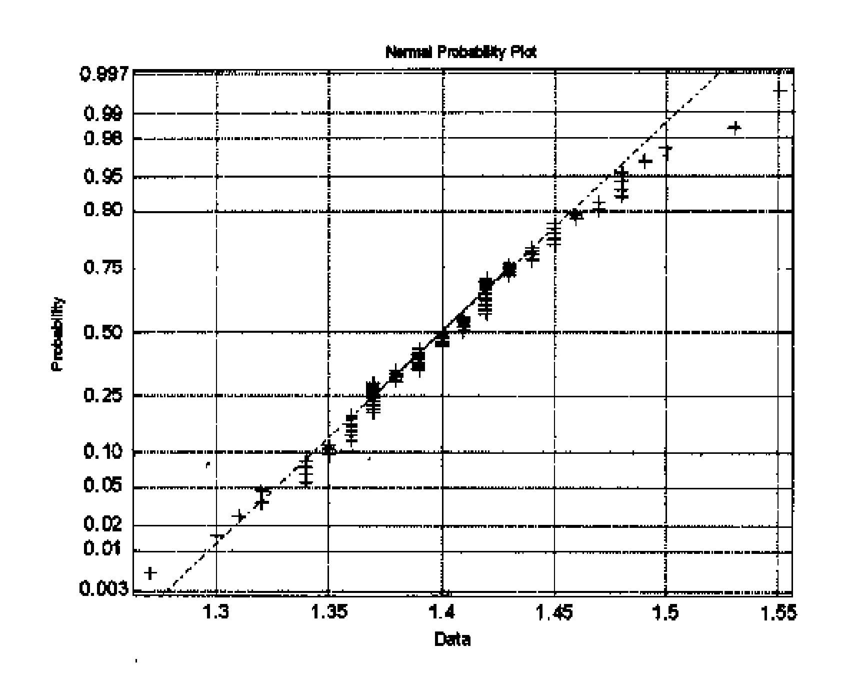


图 4.1　正态概率纸检验图

```
L_h =
      1
L_p =
      0.0451
```

③ Jarque-Bera 检验法.

```
[J_h,J_p] = jbtest(wnlxd,0.10)
```

上述指令的运行结果是:

```
J_h =
      0
J_p =
      0.3738
```

从正态概率纸检验图 4.1 可以看出, 100 个样本数据的$(x_i, 100F_n(x_i))$点列在一直线附近, 故可认为维尼纶纤度数据来自正态分布. 从图 4.1 中可以粗略地估计出维尼纶纤度的均值约为 1.4, 标准差约为 $1.45-1.4=0.05$.

Jarque-Bera 检验法的结论是接受维尼纶纤度服从正态分布的假设.

值得注意的是, Lilliefors 检验法得到的结论是拒绝维尼纶纤度服从正态分布的假设, 这是由于样本数据的标准化变换, 使得该方法对异常数据(极端数据)反应敏感. 其实, 若注意到第 99 个数据 $x_{99}=1.55$ 是 wnlxd 数据集中的最大值, 从正态概率纸检验的图形中可以看出这个最大值过于偏离直线 $y=\dfrac{1}{\sigma}(x-\mu)$, 所以 x_{99}是一个异常数据. 若

从 wnlxd 数据集中删除这个数据，重新进行检验，如下所示.

```
wnlxd(99) = [];
[h,p] = lillietest(wnlxd,0.10)
```

上述指令的运行结果是：

```
h =
    0
p =
   0.1136
```

结果表明，剩余的 99 个维尼纶纤度数据是来自正态分布的，与另外两种检验方法的结论一致.

习题 4

1. 某砖厂生产的红砖的抗断强度 X(单位：$\times 10^5$Pa)服从正态分布，设方差 $\sigma^2=1.21$，从产品中随机地抽取 6 块，测得抗断强度为 32.66，29.86，31.74，30.15，32.88，31.05. 试检验这批红砖的平均抗断强度是否为 32.50×10^5Pa？($\alpha=0.05$)

2. 某食品厂用自动装罐机装罐头食品，每罐标准质量为 500g，现从某天生产的罐头中随机抽取 9 罐，其质量(单位：g)分别为 510，505，498，503，492，502，497，506，495，假定罐头质量服从正态分布. 问：机器工作是否正常；能否认为这批罐头质量的方差为 5.5^2？($\alpha=0.05$)

3. 要比较甲、乙两种轮胎的耐磨性，现从甲、乙两种轮胎中各取 8 个，再各取一个组成一对，随机选取 8 架飞机，8 对轮胎磨损量(单位：mg)数据见表 4.12.

表 4.12

x_i(甲)	4900	5220	5500	6020	6340	7660	8650	4870
y_i(乙)	4930	4900	5140	5700	6110	6880	7930	5010

试问这两种轮胎的耐磨性有无显著差异($\alpha=0.05$)？假定甲、乙两种轮胎的磨损量分别满足 $X\sim N(\mu_1,\sigma^2)$，$Y\sim N(\mu_2,\sigma^2)$，且两个样本相互独立.

4. 生产工序中的方差是工序质量的一个重要指标. 通常，较大的方差表明具有通过寻求减小工序方差的途径来改进质量的机会.《质量管理》杂志上刊载了有关两部机器生产的袋装质量数据(以 g 为单位). 进行统计检验以确定两部机器所装袋质量的方差是否有显著差异. 取显著性水平为 0.05. 你有何结论？哪部机器有更大的改进质量的机会？假定总体服从正态分布. 两部机器所装袋质量的数据如下.

机器 1：2.95，3.45，3.50，3.75，3.48，3.26，3.33，3.20，3.16，3.20，3.22，3.38，3.90，3.36，3.25，3.28，3.20，3.22，2.98，3.45，3.70，3.34，3.18，3.35，

3.12；

机器2：3.22，3.30，3.34，3.28，3.29，3.25，3.30，3.27，3.38，3.34，3.35，3.19，3.35，3.05，3.36，3.28，3.30，3.28，3.30，3.20，3.16.

5. 某商场经理想研究家电部与服装部的每天销售额(单位：万元)的变动量是否相同，为此他收集了一周时间的每日销售额，数据如下.

家电部：5480，3500，6302，2100，3985，8670，7850；

服装部：2300，2016，2872，2559，4100，4320，4862.

假设每日销售额总体服从正态分布，试在 $\alpha=0.05$ 的显著性水平下检验两个部门的日销售额的方差是否相同.

6. 检查一本书的100页，记录各页印刷错误的个数，其结果见表4.13.

表4.13

错误个数	0	1	2	3	4	5	6	7及以上
含错误个数的页数	36	40	19	2	0	2	1	0

问能否认为一页的印刷错误个数服从泊松分布.（$\alpha=0.05$）

7. 检验下列数据是否来自正态分布($\alpha=0.05$)：

66，72，32，78，81，76，57，79，65，70，77，73，90，93，

71，74，61，86，90，90，93，77，76，66，57，81，51，65.

8. 15名新生的入学考试成绩如下：

481，620，642，515，740，525，540，598，562，395，615，596，618，584，580，

用Lilliefors检验来检验其正态性.

9. 试检验下面两组数据是否服从相同的分布？

A：8.655，10.019，9.880，8.797，9.071，9.071；

B：8.726，8.371，9.131，8.946，7.436，8.000，7.332，8.097，6.805.

10. 为了比较两种不同规格灯丝制造的灯泡使用寿命(单位：h)，分别从甲、乙两批灯泡中随机地抽取若干个灯泡进行寿命试验，测得数据如下.

甲：1420，1450，1425，1470，1465，1480；

乙：1425，1445，1410，1420，1415.

试判断这两种灯泡使用寿命是否有明显的差异.

11. 2005年“新浪”网络调查的一个问题是：“在过去的12个月中，当你公务旅行时，你最常买何种机票？”得到的数据见表4.14. 取 $\alpha=0.05$，检验航班类型与机票类型的独立性，你有何结论？

表 4.14

	航班类型	
票别	国内航班	国际航班
一等舱	29	22
商业/行政舱	95	121
企业经济舱/二等舱	518	135

12. 某企业新近推出的产品有四种款式. 欲了解不同地区顾客与新产品的不同款式是否有关，随机从三个地区抽取了 460 位顾客进行调查，获得资料见表 4.15.

表 4.15

地区＼款式	一	二	三	四	合　计
甲	25	14	20	22	81
乙	38	33	39	26	136
丙	81	46	49	67	243
合　计	144	93	108	115	460

检验不同地区与新产品的款式是否有关，即检验两者之间的关系是否相互独立. ($\alpha=0.05$)

第5章　方差分析

方差分析是重要的、应用广泛的实验数据统计分析方法，其实质是检验多个变量均值的一致性．由于检验的统计推断是通过讨论实验数据的变异性以及变异的来源作出的，而统计分析刻画数据变异性的基本统计量是样本方差，因此，习惯上称这种多变量均值一致性的假设检验为方差分析．本章介绍方差分析的基本概念、单因素方差分析和双因素方差分析方法．

5.1　方差分析概述

在实际问题的研究过程中，影响一事物的因素往往很多．例如在某化工生产中，原料成分、原料剂量、催化剂、反应时间、机器设备及操作人员等因素对产品的质量和数量都有可能产生影响．通常称试验所考查的事项(如产品的质量、数量)为**实验指标**或**响应变量**，称影响试验指标的因素(如原料成分、原料剂量、催化剂、反应时间、机器设备及操作人员等)为**试验因素**或**因子**．

试验因子对实验指标所产生的影响有大小、主次之分．在实际的试验中，人们总是控制那些次要因子使之尽可能地不发生变化，而对那些主要因子尝试不同的处理方式(置同一个因子于不同的状态)，以考查它们对实验指标的影响．例如，根据实际情况，在原料成分、原料剂量、机器设备及操作人员等因于基本保持一致的条件下，主要考查催化剂和反应时间对产品的数量指标的影响，因此选择了3种不同的催化剂(3种状态)、4种不同的反应时间(4种状态)等．通常称因子所处的状态为**因子水平**或**处理**．

实验的目的就是判断在因子的不同处理下响应变量是否有差异，以及因子最优处理是哪一种．在实验数据的统计分析中，回答这一类问题的基本方法就是比较每一种处理下响应变量的均值是否相等．在此例中，由于考虑催化剂和反应时间两个不同因子，而因子的各种处理的搭配有$3\times4=12$种，因此，产品数量这一响应变量分割为12个具体(不同处理下)的变量，若这12个变量的均值不相等，则说明催化剂和反应时间两个因子对实验的结果是有影响的．

为方便起见，今后用大写字母A，B，C等表示因子，用大写字母加下标表示该因子的水平，如因子A的水平用A_1，A_2，…表示．

为方便说明方差分析的基本思想与方法，下而考查一个简单的、易于理解的例子．

【例5.1】　一位英语教师想检查三种不同教学方法的效果，为此随机选取24名学

生并把他们分成3组，相应地用3种方法教学．一段时间后，这位教师对这24名学生进行统考，统考成绩见表5.1．试问在0.05显著性水平下，这三种教学方法有无显著性差异？

表5.1　　英语成绩表

方法	学习成绩								
A_1	73	66	89	82	43	80	63		
A_2	88	78	91	76	85	84	80	96	
A_3	68	79	71	71	87	68	59	76	80

表5.1中，A_1，A_2，A_3 是这位英语教师采用的不同教学方法，各有其侧重点．我们的目的是判断不同教学方法对英语学习成绩是否有显著影响．若有影响，哪一种教学方法好？

此例中仅有一个因子(教学方法)对实验指标或响应变量(英语成绩)可能产生影响，而因子有三种不同的处理(三种教学方法)．所以，这是一个因子的三种处理的比较问题．在进行统计分析时，将不同处理下学生的英语成绩看做三个不同的变量，仍可用 A_1, A_2, A_3 表示，并且分别记录实验数据(每一名学生的考试成绩见表5.1)，通常假定每一个变量服从方差相等的正态分布．

容易理解，不同的教学方法下学生的英语成绩可能是不同的；在同一种方法下，不同学生的英语成绩也可能是不同的．也就是说，实验数据是有差异的，而差异可能是由因子的不同处理(三种不同的教学方法)引起的，这种差异称为实验数据的**条件误差**；可能是由随机因素(不可控制或不可预知的因素，如考试时的环境、时间对学生的影响)引起的，这种差异称为实验数据的**随机误差**或**实验误差**．方差分析的主要任务就是推断在因子的不同处理下响应变量的均值(三种不同教学方法下学生的英语平均成绩)是否一致，而进行推断的基本思想就是分析实验数据的差异来源．在后面的讨论中可以看到，其中关键性的想法是考查实验数据的偏差平方和，并设想将数据总的偏差平方和按照产生的原因分解成

$$\text{总偏差平方和} = \text{条件误差平方和} + \text{随机误差平方和},$$

然后进一步比较这两种偏差平方和的大小，按照一定的统计假设检验的规则确定总的差异(总偏差平方和)究竟是由条件误差(因子的不同处理引起的偏差平方和)还是随机误差(随机因素引起的偏差平方和)决定的．如果实验数据的差异是由条件误差决定的，则说明在因子的不同处理下响应变量的均值是不同的；如果差异不是由条件误差决定的，则在因子的不同处理下响应变量的均值应当是一致的．

5.2 单因子方差分析

5.2.1 单因子试验的统计模型及检验方法

5.2.1.1 统计模型

例 5.1 中所考查的因子只有一个，称其为**单因子试验**. 通常在单因子试验中，设因子 A 有 r 个水平 $A_1, A_2, \cdots, A_r$(即试验中有 r 个处理)，在每一水平下考查的指标可以看成一个变量. 现有 r 个水平，故有 r 个变量. 为简化起见，需要给出若干假定，把所要回答的问题归结为一个统计问题，然后设法解决它. 假定：

① 每一变量均服从正态分布；

② 每一变量的方差相同；

③ 从 r 个变量抽取的样本相互独立.

我们要比较各个变量的均值是否一致，设第 i 个变量的均值为 μ_i，那么就要检验如下假设：

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r,$$

其备择假设为

$$H_1: \mu_1, \mu_2, \cdots, \mu_r \text{ 不全相同}.$$

通常 H_1 可以省略不写.

当 H_0 为真时，称因子 A 的各水平间无显著差异，简称**因子 A 不显著**(此时在例 5.1 中得出不同的教学方法对英语学习成绩没有显著影响)；反之，当 H_0 不真时，各 μ_i 不全相同，这时称因子 A 的各水平间有显著差异，简称**因子 A 显著**.

用于检验假设 H_0 的统计方法称为**方差分析法**，其实质是检验若干个具有相同方差的正态变量的均值是否相等的一种统计方法. 在所考虑的因子仅有一个的场合，称为**单因子方差分析**.

为检验假设 H_0，需要对每一变量抽取样本. 这些样本可以通过试验或某种观察获得. 各样本间还是相互独立的. 为方便起见，本章对样本及其观察值都用同一符号 y 加下标表示，其含义可从上下文理解. 设第 i 个变量对应容量为 m_i 的样本 $y_{i1}, \cdots, y_{im_i}$ $(i=1,2,\cdots,r)$.

在 A_i 水平下获得的 y_{ij} 与 μ_i 不会总是一致的，如例 5.1 中教学方法 A_1 下学生的成绩也不完全相同. 记

$$\varepsilon_{ij} = y_{ij} - \mu_i,$$

称 ε_{ij} 为**随机误差**，从而有

$$y_{ij}=\mu_i+\varepsilon_{ij},$$

称上式为 y_{ij} 的**数据结构式**，即来自均值为 μ_i 的变量观察值 y_{ij} 可看成是由其均值 μ_i 与随机误差 ε_{ij} 叠加而产生的．在假定 A_i 的指标 y_{ij} 服从 $N(\mu_i,\ \sigma^2)$ 分布时，则有 $\varepsilon_{ij}\sim N(0,\sigma^2)$．

综上，有单因子方差分析的统计模型：假定

$$\left.\begin{array}{l} y_{ij}=\mu_i+\varepsilon_{ij}, \\ \varepsilon_{ij}\sim N(0,\sigma^2)\text{且相互独立}, \end{array}\quad (i=1,2,\cdots,r;\ j=1,2,\cdots,m_i)\right\} \qquad (*)$$

检验假设 $H_0:\mu_1=\mu_2=\cdots=\mu_r$．

为了能更仔细地描述数据，常在方差分析模型中引入一般平均与效应的概念．称诸 μ_i 的加权平均

$$\mu=\frac{1}{n}\sum_{i=1}^{r}m_i\mu_i$$

为一般平均，其中 $n=\sum_{i=1}^{r}m_i$．称

$$a_i=\mu_i-\mu\quad(i=1,2,\cdots,r)$$

为因子 A 第 i 水平的**主效应**，也简称为 A_i 的**效应**．容易看出，效应间有如下关系式：

$$\sum_{i=1}^{r}m_i a_i=0.$$

在上述记号下，有

$$\mu_i=\mu+a_i.$$

这表明第 i 个总体的均值是一般平均与其效应的叠加．此时单因子方差分析的统计模型可改写成

$$\begin{cases} y_{ij}=\mu+a_i+\varepsilon_{ij}, \\ \sum_{i=1}^{r}m_i a_i=0, & (i=1,2,\cdots,r;\ j=1,2,\cdots,m_i) \\ \varepsilon_{ij}\sim N(0,\sigma^2)\text{ 且相互独立}, \end{cases}$$

它由数据结构式、关于效应的约束条件及关于误差的假定三部分组成．在上述模型下，所要检验的假设可改写成 $H_0:a_1=a_2=\cdots=a_r=0$．

5.2.1.2　检验方法

在单因子方差分析中，通常将所得数据列成如表 5.2 所示的形式．

表 5.2 中，各 y_{ij} 是有差异的，我们从考查数据间的差异着手来给出检验方法．

造成各 y_{ij}间差异的原因可能有两个：一个可能是假设 H_0 不真，即各水平下变量均值 μ_i(或水平效应 a_i)不同，因此从各变量获得的样本观测值也有差异；另一可能是 H_0 为真，差异是由随机误差引起的.

表 5.2　　单因子方差分析数据结构表

因子水平	试验数据			
A_1	y_{11}	y_{12}	…	y_{1m_1}
A_2	y_{21}	y_{22}	…	y_{2m_2}
⋮	⋮	⋮		⋮
A_r	y_{r1}	y_{r2}	…	y_{rm_r}

为使这些差异的大小能定量表示出来，先引入如下若干记号.

把 A_i 水平下试验数据和记为$y_{i.}=\sum_{j=1}^{m_i}y_{ij}$，其平均值记为 $\bar{y}_{i.}=\frac{1}{m_i}y_{i.}$，由 y_{ij}的数据结构式可知，$y_{i.}$具有如下结构式：

$$\bar{y}_{i.}=\mu_i+\bar{\varepsilon}_{i.},$$

其中 $\bar{\varepsilon}_{i.}=\frac{1}{m_i}\sum_{j=1}^{m_i}\varepsilon_{ij}$.

把所有数据之和记为 $y_{..}=\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}$，其平均值记为 $\bar{y}=\frac{y_{..}}{n}$，$\bar{y}$ 具有如下结构式：

$$\bar{y}=\mu+\bar{\varepsilon},$$

其中 $\bar{\varepsilon}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m_i}\varepsilon_{ij}$. 由于

$$y_{ij}-\bar{y}=(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}),$$

其中 $y_{ij}-\bar{y}_{i.}$称为**组内偏差**，仅反映随机误差：

$$y_{ij}-\bar{y}_{i.}=(\mu_i+\varepsilon_{ij})-(\mu_i+\bar{\varepsilon}_{i.})=\varepsilon_{ij}-\bar{\varepsilon}_{i.}.$$

而 $\bar{y}_{i.}-\bar{y}$ 称为**组间偏差**，除了反映随机误差外，还反映了第 i 个水平效应：

$$\bar{y}_{i.}-\bar{y}=(\mu_i+\bar{\varepsilon}_{i.})-(\mu+\bar{\varepsilon})=a_i+\bar{\varepsilon}_{i.}-\bar{\varepsilon}.$$

各 y_{ij}间总的差异大小可用**总偏差平方和** SST 表示：

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y})^2.$$

由随机误差引起的数据间的差异可以用组内偏差平方和表示. 由于组内偏差仅反映随机误差，故也把组内偏差平方和称为**误差偏差平方和**，记为 SSE：

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2.$$

由于组间偏差除了反映随机误差外，还反映了效应间的差异，故由效应不同引起的数据差异可用组间偏差平方和表示，也称为**因子 A 的偏差平方和**，记为 SSA：

$$SSA = \sum_{i=1}^{r} m_i(\bar{y}_{i\cdot} - \bar{y})^2.$$

这里，每一项乘上 m_i 是因为第 i 水平有 m_i 个实验数据.

定理 5.1（平方和分解定理 1） $SST = SSA + SSE$.

事实上

$$\begin{aligned} SST &= \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{m_i}(\bar{y}_{i\cdot} - \bar{y})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}) \\ &= SSE + SSA. \end{aligned}$$

由于 $\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot}) = 0$，故上述第三项为 0.

由模型(*)可知各 ε_{ij} 相互独立，且 $\varepsilon_{ij} \sim N(0, \sigma^2)$ $(i = 1, 2, \cdots, r;\ j = 1, 2, \cdots, m_i)$，故

$$\bar{\varepsilon}_{i\cdot} \sim N\left(0, \frac{\sigma^2}{m_i}\right) \quad (i = 1, 2, \cdots, r),$$

$$\bar{\varepsilon} \sim N\left(0, \frac{\sigma^2}{n}\right).$$

由于

$$\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})^2 = \frac{1}{\sigma^2}\sum_{j=1}^{m_i}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2 \sim \chi^2(m_i - 1),$$

又由 χ^2 分布的可加性可知

$$\frac{SSE}{\sigma^2} = \sum_{i=1}^{r}\left[\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})^2\right] \sim \chi^2\left(\sum_{i=1}^{r}(m_i - 1)\right) = \chi^2(n - r).$$

由 χ^2 分布的性质知

$$E\left(\frac{SSE}{\sigma^2}\right) = n - r,$$

即

$$E(SSE) = (n - r)\sigma^2.$$

由于

$$SSA = \sum_{i=1}^{r} m_i(\bar{y}_{i\cdot} - \bar{y})^2 = \sum_{i=1}^{r} m_i(a_i + \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2$$

$$= \sum_{i=1}^{r} m_i a_i^2 + \sum_{i=1}^{r} m_i \bar{\varepsilon}_{i\cdot}^2 - n\bar{\varepsilon}^2 + 2\sum_{i=1}^{r} m_i a_i(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}),$$

又由 $E(\bar{\varepsilon}_{i\cdot})=0$, $E(\bar{\varepsilon})=0$, 故

$$E(SSA) = \sum_{i=1}^{r} m_i a_i^2 + \sum_{i=1}^{r} m_i E(\bar{\varepsilon}_{i\cdot}^2) - nE(\bar{\varepsilon}^2) = \sum_{i=1}^{r} m_i a_i^2 + \sum_{i=1}^{r} m_i \frac{\sigma^2}{m_i} - n \cdot \frac{\sigma^2}{n}$$

$$= \sum_{i=1}^{r} m_i a_i^2 + (r-1)\sigma^2.$$

从上面的分析过程中可得如下定理.

定理 5.2 (平方和期望定理) 在一个因素的方差分析模型中, 有

$$E(SSE) = (n-r)\sigma^2,$$

$$E(SSA) = \sum_{i=1}^{r} m_i a_i^2 + (r-1)\sigma^2.$$

定理 5.3 (误差偏差平方和分布定理) 在一个因素的方差分析模型中, 有

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-r).$$

定理 5.4 (因子 A 的偏差平方和分布定理) 在一个因素的方差分析模型中, 当假设 H_0 为真时, 有

$$E\left(\frac{SSA}{r-1}\right) = \sigma^2,$$

$$\frac{SSA}{\sigma^2} \sim \chi^2(r-1),$$

SSA 与 SSE 相互独立, 且 $F = \dfrac{SSA/(r-1)}{SSE/(n-r)} \sim F(r-1, n-r)$.

定理 5.2, 5.3, 5.4 的证明参见文献[2].

因此可采用统计量 F 来检验假设 H_0. 当 H_0 不真时, 分子的均值要比分母的均值大, 因而取如下拒绝域

$$W = \{F \geqslant c\}$$

是合理的. 对给定的显著性水平 α, c 应满足

$$P\{F \geqslant c\} = \alpha,$$

当取 $c = F_{1-\alpha}(r-1, n-r)$时, 便有 $P\{F \geqslant c\} = \alpha$, 故得拒绝域为

$$W = \{F \geqslant F_{1-\alpha}(r-1, n-r)\}.$$

通常把以上求统计量的计算列成一张表格, 称为方差分析表(见表 5.3), 相应的 χ^2 分布中的自由度也列于表中, 偏差平方和与自由度的比称为均方和.

表 5.3　　单因子方差分析表

偏差来源	偏差平方和	自由度	均方和	F 值
A	SSA	$f_A = r-1$	$V_A = SSA/f_A$	$F = V_A/V_E$
E	SSE	$f_E = n-r$	$V_E = SSE/f_E$	
T	SST	$f_T = n-1$		

综上，作单因子方差分析的步骤如下.

① 依次列出第 i 个变量($i=1,2,\cdots,r$)对应容量为 m_i 的样本 $y_{i1},\cdots,y_{im_i}$，确定试验中因子的水平数 r、各水平下的样本容量 m_i、数据总数 $n=\sum_{i=1}^{r}m_i$，同时明确显著性水平 α.

② 计算各水平下数据和 $y_{i\cdot}=\sum_{j=1}^{m_i}y_{ij}\ (i=1,2,\cdots,r)$及总和 $y_{\cdot\cdot}=\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}$，计算各数据 y_{ij}平方之和 $\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}^2$，在此基础上计算 $\sum_{i=1}^{r}\frac{y_{i\cdot}^2}{m_i}$，$\frac{y_{\cdot\cdot}^2}{n}$.

③ 利用步骤②中的结果，计算 SST，SSA 和 SSE. 其中

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}^2-\frac{y_{\cdot\cdot}^2}{n},$$

$$SSA=\sum_{i=1}^{r}m_i(\bar{y}_{i\cdot}-\bar{y})^2=\sum_{i=1}^{r}\frac{y_{i\cdot}^2}{m_i}-\frac{y_{\cdot\cdot}^2}{n},$$

$$SSE=SST-SSA.$$

④ 确定自由度 $f_A=r-1$ 和 $f_E=n-r$，计算各类均方和 $V_A=SSA/f_A$ 和 $V_E=SSE/f_E$，求出检验用统计值 $F=V_A/V_E$，即得到了单因子方差分析表中的各项内容.

⑤ 求出临界值 $F_{1-\alpha}(f_A,f_E)$，确定拒绝域 $W=\{F\geqslant F_{1-\alpha}(f_A,f_E)\}$. 若 $F\in W$，则作出拒绝原假设 H_0 的结论；否则，接受 H_0.

或者由最小显著性概率 p 作出检验决策，当 $p<\alpha$ 时拒绝原假设.

对于例 5.1，所谓方差分析，即检验如下假设：$H_0:\mu_1=\mu_2=\mu_3$，其中 $\mu_i(i=1,2,3)$是第 i 个变量的均值. 按照上述步骤，具体的检验过程可由如下 MATLAB 指令集完成.

MATLAB 数据处理(1)

```
clear
y=[73,66,89,82,43,80,63,88,78,91,76,85,94,80,96,68,79,71,71,87,68,59,
76,80];
r=3;
m1=7;m2=8;m3=9; % 各总体的样本容量
```

```
n = m1 + m2 + m3;
alpha = 0.05;
y1_ = sum(y(1:m1)); % 第一种教学方法下学生的成绩之和
y2_ = sum(y((m1 + 1):(m1 + m2))); % 第二种教学方法下学生的成绩之和
y3_ = sum(y((m1 + m2 + 1):n)); % 第三种教学方法下学生的成绩之和
y_ = sum(y); % 各学生成绩之和
yy = sum(y.^2); % 各学生成绩平方之和
g = y1_^2/m1 + y2_^2/m2 + y3_^2/m3;
SST = yy - y_^2/n; % 总的偏差平方和
SSA = g - y_^2/n; % 因子的偏差平方和
SSE = SST - SSA; % 误差平方和
g1 = SSA/(r - 1); % 偏差均方和
g2 = SSE/(n - r); % 误差均方和
FEST = g1/g2; % 由样本计算出的 F 值
FLJ = finv(1 - alpha,r - 1,n - r); % 应用 MATLAB 统计工具箱中 finv 函数求得临
界值
p = 1 - fcdf(FEST,r - 1,n - r);
if FEST>FLJ
h = 1;
else
h = 0;
end
alpha, h, p, FEST, FLJ
```

上述指令的运行结果是：

```
alpha =
     0.0500
h =
     1
p =
     0.0211
FEST =
     4.6638
FLJ =
     3.4668
```

计算结果表明，在 0.05 显著性水平下，h=1、p<alpha 拒绝原假设，即认为三种教学方法有显著性差异.

5.2.2 效应与误差方差的估计

5.2.2.1 效应与误差方差的点估计

由模型(*)知各 y_{ij} 相互独立，且 $y_{ij}\sim N(\mu+a_i,\sigma^2)$，因而可用极大似然法求出各效应与 σ^2 的估计. 不难证明如下定理.

定理 5.5 (效应与误差方差的点估计定理)

$$\hat{\mu}=\bar{y},\quad \hat{\mu}_i=\bar{y}_{i\cdot},\quad \hat{a}_i=\bar{y}_{i\cdot}-\bar{y}\quad (i=1,2,\cdots,r),$$

$$\hat{\sigma}_M^2=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{SSE}{n},$$

$$\sigma^2 \text{ 的无偏估计是 } \hat{\sigma}^2=\frac{SSE}{n-r}.$$

证明参见文献[1].

5.2.2.2 μ_i 的置信水平为 $1-\alpha$ 的置信区间

利用枢轴量法，可以构造 μ_i 的置信区间.

从 μ_i 的点估计 $\bar{y}_{i\cdot}$ 出发，由于前已证明 $\bar{y}_{i\cdot}\sim N\left(\mu_i,\frac{\sigma^2}{m_i}\right)$，又 $\frac{SSE}{\sigma^2}\sim\chi^2(f_E)$，这里 $f_E=n-r$，且 $\bar{y}_{i\cdot}$ 与 SSE 独立，因而可以构造一个服从 t 分布的枢轴量

$$t_i=\frac{\dfrac{\bar{y}_{i\cdot}-\mu_i}{\dfrac{\sigma}{\sqrt{m_i}}}}{\sqrt{\dfrac{\dfrac{SSE}{\sigma^2}}{f_E}}}=\frac{\bar{y}_{i\cdot}-\mu}{\dfrac{\hat{\sigma}}{\sqrt{m_i}}}\sim t(f_E),$$

因而从

$$P\left\{\left|t_i\right|\leqslant t_{1-\frac{\alpha}{2}}(f_E)\right\}=1-\alpha$$

可得 μ_i 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\bar{y}_{i\cdot}-t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m_i}},\ \bar{y}_{i\cdot}+t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m_i}}\right),$$

这里 $\hat{\sigma}=\sqrt{\dfrac{SSE}{f_E}}$.

【例 5.2】 求例 5.1 中每一种教学方法下学生平均英语成绩的点估计和置信水平为 0.95 的置信区间.

按照本小节的定理和结论，利用 MATLAB 进行计算，具体过程如下.

MATLAB 数据处理(2)

```
clear
alpha = 0.05;
m1 = 7; m2 = 8; m3 = 9;
n = m1 + m2 + m3;
r = 3;
fE = n - r;
y1_ = 496; % 引用 MATLAB 数据处理(1)中结果, 下同
y2_ = 688;
y3_ = 659;
MU1 = y1_/m1 % 第一种教学方法下学生平均英语成绩的点估计
MU2 = y2_/m2 % 第二种教学方法下学生平均英语成绩的点估计
MU3 = y1_/m3 % 第三种教学方法下学生平均英语成绩的点估计
T = tinv(1 - alpha/2, fE);
SSE = 2.3404e + 003; % 引用 MATLAB 数据处理(1)中结果
SIGMA = sqrt(SSE/(n - r)); % 英语成绩标准差的无偏估计
a = [MU1 - T * SIGMA/sqrt(m1), MU1 + T * SIGMA/sqrt(m1)];
b = [MU2 - T * SIGMA/sqrt(m2), MU2 + T * SIGMA/sqrt(m2)];
c = [MU3 - T * SIGMA/sqrt(m3), MU3 + T * SIGMA/sqrt(m3)];
a, b, c % 三种教学方法下平均英语成绩的置信区间
```

上述指令的运行结果是：

```
MU1 =
   70.8571
MU2 =
   86
MU3 =
   55.1111
a =
   62.5592     79.1551
b =
   78.2380     93.7620
```

```
c =

    47.7930    62.4292
```

计算结果表明，三种教学方法下学生的平均英语成绩分别为 70.8571，86，55.1111；95% 的置信区间分别为 [62.5592，79.1551]，[78.2380，93.7620]，[47.7930，62.4292].

5.2.3 重复数相同的方差分析

当在因子 A 的每一水平下重复试验次数相同，即当 $m_1 = m_2 = \cdots = m_r$ 时，上述一些表达式可以简化．若记每一水平下重复次数为 m，则

效应约束条件可简化为 $\sum_{i=1}^{r} a_i = 0$；

SSA 的计算公式可简化为 $SSA = \frac{1}{m}\sum_{i=1}^{r} y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{n}$；

μ_i 的置信水平为 $1-\alpha$ 的置信区间可改为

$$\left(\bar{y}_{i\cdot} - t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m}},\ \bar{y}_{i\cdot} + t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m}}\right).$$

其他一切都不变．对于重复数相同的单因子方差分析，MATLAB 提供了命令函数 anoval 来处理单因素方差分析的问题．命令 anoval 主要是比较多组数据的均值，然后返回这些均值相等的概率，从而判断这一因素是否对试验指标有显著影响．调用方法：

```
[p, anovatab, stats] = anova1(X, group, 'displayopt')
```

其中，输入参数 X 表示 r 变量的 m 个样本观测值的 $m \times r$ 矩阵．group 是与 X 对应的表示 r 变量名字或意义字符串数组，通常缺省使用．引用参数 displayopt 有两个状态 on 和 off，分别表示显示和隐藏方差分析表图形和 box 图．输出参数 p 为 X 的各列均值相等的最小显著性概率，p 的值越小，则质疑原假设，表示这个因素对随机变量的影响是显著的．anovatab 和 stats 分别返回方差分析表和一个附加的统计数据结构，可以缺省．

【例 5.3】 某钢厂检查一月上旬的五天中生产的钢锭质量，结果见表 5.4(单位：kg).

表 5.4

日　期	质　　量			
1	5500	5800	5740	5710
2	5440	5680	5240	5600
4	5400	5410	5430	5400
9	5640	5700	5660	5700
10	5610	5700	5610	5400

试检验不同日期生产的钢锭有无显著差异？($\alpha=0.05$)

分析 我们把不同日期生产的钢锭质量分别看做一个变量. 检验它们的平均质量是否有明显差异相当于要比较五个变量的均值是否一致. 假定：①五个变量均服从正态分布；②每一变量的方差相同；③从五个变量抽取的样本相互独立. 采用方差分析法来检验不同日期生产的钢锭质量是否有明显差异.

设第 i 个变量的均值为 μ_i，假设不同日期生产的钢锭平均质量无显著差异. 那么就要检验如下假设：

$$H_0: \mu_1=\mu_2=\mu_3=\mu_4=\mu_5.$$

具体见以下解题过程.

MATLAB 数据处理

```
clear
A1 = [5500, 5800, 5740, 5710]';
A2 = [5440, 5680, 5240, 5600]';
A3 = [5400, 5410, 5430, 5400]';
A4 = [5640, 5700, 5660, 5700]';
A5 = [5610, 5700, 5610, 5400]';
X = [A1, A2, A3, A4, A5];
[p, anovatab, stats] = anova1(X, [], 'on')
```

上述指令的运行结果见图 5.1 及：

```
p =
    0.0220
anovatab =
    'Source'    'SS'       'df'    'MS'               'F'         'Prob>F'
    'Columns'   [227680]   [ 4]    [    56920]        [3.9496]    [0.0220]
    'Error'     [216175]   [15]    [1.4412e+004]      []          []
    'Total'     [443855]   [19]    []                 []          []
stats =
        gnames: [5x1 char]
             n: [4 4 4 4 4]
        source: 'anova1'
         means: [5.6875e+003 5490 5410 5675 5580]
            df: 15
             s: 120.0486
```

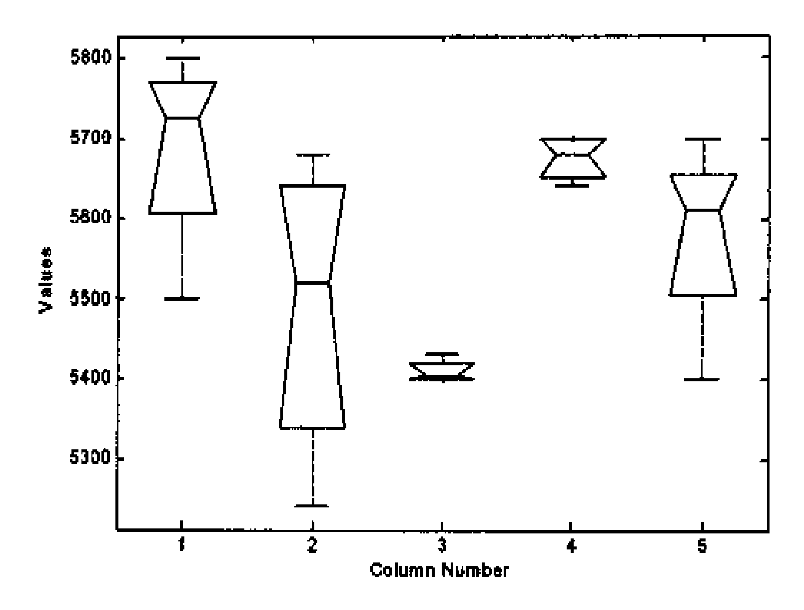


图 5.1　五天生产钢锭质量的 box 图

结果表明：①返回值 p=0.0220<0.05，认为不同日期生产的钢锭平均质量有显著差异. ②方差分析表(anovatab)中有 6 列，第 1 列声明 X 中可变化性的来源；第 2 列显示平方和；第 3 列显示与每一种可变性有关的自由度；第 4 列显示第 2 列数据与第 3 列数据的比值；第 5 列显示 F 统计量数值，是第 4 列数据的比值；第 6 列检验的最小显著性概率，即第一输出参数值. ③stats 返回的附加统计数据结构中 means 一行给出了各日生产的钢锭平均质量的点估计. ④从方差分析 box 图容易看出不同日期生产的钢锭平均质量之间的直观差异.

5.2.4　多重比较

若检验结果拒绝了 H_0，进一步分析哪些水平之间的差异是显著的、哪些水平对实验结果的影响最大、哪些水平次之，这在实际应用中往往是很重要的. 此项工作通常称为均值的多重比较.

对任意两个水平均值之间有无显著差异进行多重比较，即同时检验以下 $\binom{r}{2}$ 个假设：

$$H_0^{ij}:\mu_i=\mu_j,\quad H_1^{ij}:\mu_i\neq\mu_j\quad(i<j;\ i,j=1,2,\cdots,r).$$

检验的统计量为

$$t=\frac{(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})}{\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}},$$

其中 $s^2=\frac{SSE}{n-r}$. 对于 H_0^{ij} 的检验水平 α'，当 $|t|>t_{1-\frac{\alpha'}{2}}(n-r)$时拒绝 H_0^{ij}. 或等价地，

当置信度为 $100(1-\alpha')\%$ 的 $\mu_i-\mu_j$ 置信区间

$$t=(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})\pm t_{1-\frac{\alpha'}{2}}(n-r)\cdot s\cdot\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

不包含 0 时拒绝 H_0^{ij}，从而拒绝 H_0.

由于多重比较所进行的一系列检验均构成对于假设的检验，因此要使得所有检验总的犯第一类错误的概率不超过给定的 α，就需要选取适当的 α'. 检验 H_0 和检验 H_0^{ij} 的交 $\bigcap\limits_{1\leqslant i<j\leqslant r}H_0^{ij}$ 等价：当所有的 H_0^{ij} 成立时，H_0 必成立，反之亦然. 以 A_{ij}记H_0^{ij} 的拒绝域，则

$$\begin{aligned}P(\text{拒绝 } H_0|H_0)&=P(\text{至少有一个 } A_{ij}\text{发生}|H_0)\\&=P(A_{12}+A_{13}+\cdots+A_{r,\,r-1}|H_0)\\&\leqslant\sum_{1\leqslant i<j\leqslant r}P(A_{ij}\mid H_0)\\&\leqslant\sum_{1\leqslant i<j\leqslant r}P(A_{ij}\mid H_0^{ij})\leqslant\binom{r}{2}\alpha'.\end{aligned}$$

要使总的犯第一类错误的概率 $P(\text{拒绝 } H_0|H_0)\leqslant\alpha$，只要取 $\alpha'=\alpha/\binom{r}{2}$.

通过$\binom{r}{2}$个两均值比较，检验假设 H_0 的优点是它不仅可知 $\mu_1,\mu_2,\cdots,\mu_r$ 有差别，而且知道差别在哪. 但此方法计算量大，同时由于要保证总的检验水平，α'取得比较小，从而一般说来，比起直接应用方差分析增大了犯第二类错误的概率，这意味着可能会出现这样的情形：用 F 检验结果是显著的，但用两两比较却没有任何两个水平有显著差异. 下面的 LSD 方法在某种程度上可以弥补这个缺陷，但真实水平是近似的.

LSD 方法是由 R.A.*Fisher* 提出，又经过后人修正的. 方法如下：

① 给定检验水平 α，用方差分析法检验 H_0；

② 如果拒绝 H_0，则继续比较水平之间的差异，否则停止；

③ 对于水平 i,j，μ_i 与 μ_j 的最小显著差异为

$$\mathrm{LSD}_{ij}=t_{1-\frac{\alpha}{2}}(n-r)\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)};$$

④ 当 $|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\geqslant\mathrm{LSD}_{ij}$时，认为 μ_i 与 μ_j 不同.

【例 5.4】 用多重比较的方法确定例 5.1 中哪些水平之间的差异是显著的，同时确定使学生的平均英语成绩最高的那种教学方法.

分析 例 5.1 中，我们已经得出三种教学方法有显著性差异，即教学方法这一因子对学生的英语成绩是有显著影响的. 进一步分析到底哪两种教学方法对学生的成绩影响差异显著，就需要对三个变量进行多重比较了. 多重比较的方法很多，按照上面介绍的

LSD 方法，利用 MATLAB 计算如下.

MATLAB 数据处理(3)

```
clear
alpha = 0.05;
m1 = 7; m2 = 8; m3 = 9;
n = m1 + m2 + m3;
r = 3;
t = tinv(1 - alpha/2, n - r);
SSE = 2.3404e + 003; % 引用 MATLAB 数据处理(1)中结果
LSD12 = t * sqrt(SSE/(n - r)) * sqrt(1/m1 + 1/m2);
LSD13 = t * sqrt(SSE/(n - r)) * sqrt(1/m1 + 1/m3);
LSD23 = t * sqrt(SSE/(n - r)) * sqrt(1/m2 + 1/m3);
MU1 = 70.8571; % 引用 MATLAB 数据处理(2)中结果, 下同
MU2 = 86;
MU3 = 55.1111;
if abs(MU1 - MU2) > = LSD12
h(1) = 1;
else
h(1) = 0;
end
if abs(MU1 - MU3) > = LSD13
h(2) = 1;
else
h(2) = 0;
end
if abs(MU2 - MU3) > = LSD23
h(3) = 1;
else
h(3) = 0;
end
h % 结果,依次显示第 1 和 2, 1 和 3, 2 和 3 种方法下学生平均成绩差异的显著性
```

上述指令的运行结果是:

```
h =
     1     1     1
```

计算结果表明：三种教学方法对学生英语平均成绩的影响有显著差异；第二种教学方法使学生的英语平均成绩最高.

5.2.5 方差齐性检验

在单因子方差分析中，假定 r 个不同水平下的响应变量 y_i 服从 $N(\mu_i, \sigma_i^2)$ $(i=1,2,\cdots,r)$，并要求这 r 个正态变量的方差相等. 这一要求简称为方差齐性. 一般而言，实际应用中在进行方差分析之前，有两项预备性分析是不可或缺的. 一是这 r 个变量的正态性检验，检验方法在第 4 章已作介绍；另一是这 r 个正态变量的方差齐性检验，本小节扼要介绍这一问题的检验方法.

方差齐性检验的假设为

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2; \quad H_1: \sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2 \text{ 不全相等}.$$

备择假设往往略去不写.

方差齐性通常采用 Bartlett 检验方法. 下面简单介绍 Bartlett 检验的基本思路和检验统计量的构造.

设第 i 个变量抽取了容量为 m_i 的样本 $y_{i1}, y_{i2}, \cdots, y_{im_i}$，其样本方差为

$$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})^2 = \frac{Q_i}{f_i} \quad (i = 1,2,\cdots,r),$$

其中 $Q_i = \sum_{i=1}^{m_i}(y_{ij} - \bar{y}_i)$，$f_i = m_i - 1$ 分别为该变量的样本偏差平方和与自由度. 于是，随机误差均方和

$$MSSE = \frac{1}{f_E}SSE = \frac{1}{f_E}\sum_{i=1}^{r} Q_i = \sum_{i=1}^{r}\frac{f_i}{f_E}s_i^2$$

是 r 个变量样本方差 $s_i^2(i=1,2,\cdots,r)$ 的加权算术平均数. 又令

$$GMSSE = \left[\prod_{i=1}^{r}(s_i^2)^{f_i}\right]^{\frac{1}{f_E}}$$

是 r 个变量样本方差 $s_i^2(i=1,2,\cdots,r)$ 的几何平均数，$f_E = \sum_{i=1}^{r} f_i$.

由于恒有 $GMSSE \leqslant MSSE$，并且等号成立的充分必要条件是 $s_1^2 = s_2^2 = \cdots = s_r^2$，所以，诸样本方差 $s_i^2(i=1,2,\cdots,r)$ 间的差异越大，$GMSSE$ 和 $MSSE$ 的差异越大. 换句话说，当 H_0 为真时，比值 $MSSE/GMSSE$ 接近于 1. 反之，比值 $MSSE/GMSSE$ 较大时，H_0 值得怀疑. 这个结论对 $\ln(MSSE/GMSSE)$ 也成立. 于是，H_0 的拒绝域应有如下形式：

$$W = \{\ln(MSSE/GMSSE) \geqslant d\}.$$

Bartlett 证明了，在大样本条件下

$$B=\frac{f_E}{c}(\ln MSSE-\ln GMSSE)\dot{\sim}\chi^2(r-1),$$

其中 $c=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{f}{f_E}\right)$. 显然，一般情况下 $c>1$.

通常，当各个变量的样本容量 $m_i\geqslant5$ $(i=1,2,\cdots,r)$时，也可以用统计量 B 作为 H_0 的检验统计量，在显著性水平 α 下，拒绝域为

$$W=\{B\geqslant\chi^2_{1-\alpha}(r-1)\}.$$

实际计算时，检验统计量采用

$$B=\frac{1}{c}\left(f_E\ln(SSE/f_E)-\sum_{i=1}^{r}f_i\ln s_i^2\right)$$

的形式更方便一些.

【例 5.5】 对例 5.1 中三种教学方法下学生的英语成绩这三个变量作方差齐性检验.

分析 假设 $H_0:\sigma_1^2=\sigma_2^2=\sigma_3^2$，即三个变量的方差相等. 按照上述结论，分别求得例 5.1 中检验统计量 B 的值和本题的拒绝域，经过比较得出结论.

MATLAB 数据处理(4)

```
clear
y=[73,66,89,82,43,80,63,88,78,91,76,85,94,80,96,68,79,71,71,87,68,59,
76,80];
alpha =0.05;
m1=7;m2=8;m3=9;
r=3;
SSE=2.3404e+003; % 引用 MATLAB 数据处理(1)中结果
n=m1+m2+m3;
fE=n-r;
c=(1/(m1-1)+1/(m2-1)+1/(m3-1)-1/fE)/(3*(r-1))+1;
s1=var(y(1:m1)); s2=var(y((m1+1):(m1+m2))); s3=var(y((n-m3+1):
n));
chi2EST=(fE*log(SSE/fE)-(m1-1)*log(s1)-(m2-1)*log(s2)-(m3-1)*
log(s3))/c;
LJZ=chi2inv(1-alpha,r-1);
p=1-chi2cdf(chi2EST,r-1);
if chi2EST>LJZ
```

```
h = 1;
else
h = 0;
end
alpha, h, p, chi2EST, LJZ
```

上述指令的运行结果是：

```
alpha =
     0.0500
h =
     0
p =
     0.1330
chi2EST =
     4.0348
LJZ =
     5.9915
```

计算结果表明，在0.05显著性水平下，h=0、p>alpha不能拒绝原假设，即认为三种教学方法下学生的英语成绩这三个变量方差相等.

下面，对单因子方差分析的应用步骤小结如下.

① 对各个变量(不同的因子水平)的正态性进行检验(例5.1中忽略了这一步).

② 对各个变量的方差齐性进行检验(如例5.1中MATLAB数据处理(4)).

③ 当各个变量的正态性和方差齐性得到验证后，进行方差分析(如例5.1中MATLAB数据处理(1)). 在各个变量的正态性和方差齐性没有得到验证的情况下，严格地说不宜再作方差分析. 但是，有关研究表明方差分析的F统计量有较好的稳健性，即使正态性和方差齐性没有得到验证也可以进行粗略的方差分析以供参考.

④ 在方差分析拒绝各个变量均值一致的原假设后，应进行多重比较(如例5.1中MATLAB数据处理(3)).

⑤ 无论方差分析是否拒绝原假设，都应对每个变量的均值作出估计(如例5.1中MATLAB数据处理(2)).

5.3 双因子方差分析

在许多实际问题中，常常需要同时研究几个因子对实验指标的影响作用. 如在例5.1中，学生的英语成绩不仅与教学方法有关，也与其自身的努力程度等因素有关系.

为使讨论相对直观一些，结合下面的例题阐述无交互作用的双因子方差分析方法与有交互作用的双因子方差分析方法.

【例 5.6】 表 5.5 中数据是在 4 个地区种植的 3 种松树的直径(单位:cm).

表 5.5

树种	地区 1					地区 2					地区 3					地区 4				
A	23	15	26	13	21	25	20	21	16	18	21	24	24	29	19	14	11	19	20	24
B	28	22	25	19	26	30	26	26	20	28	17	27	19	23	13	17	21	18	26	23
C	18	10	12	22	13	15	21	22	14	12	16	19	25	25	22	18	12	23	22	19

试问：(1) 是否有某种树特别适合在某地区种植?

(2) 若(1)是否定的，各树种有无差别? 哪种树最好? 哪个地区最适合松树生长?

5.3.1 无交互作用的双因子方差分析

设 A 与 B 是对试验结果有影响的两个因子，相互独立. 如例 5.6 中，树种和地区便是影响松树生长的两个因子，这里我们以松树的直径大小作为判断松树生长优良的实验指标. 现在只是不知道这两个因子之间是否存在交互作用，即是否存在某个地区最适合某种松树生长. 这种情况下，应首先按照有交互作用的方差分析方法去检验因子之间交互作用的存在性. 如果根据生产实际经验或有关专业知识，知道它们之间不存在交互作用，或者它们的交互作用不显著，可以忽略不计.

首先讨论因子之间无交互作用的情形.

仅仅为分析因子 A 与因子 B 各自对实验指标的影响是否显著而设计的试验可以是无重复试验，即各种水平组合只进行一次试验，各获得一个试验数据就够了. 因子 A 有 r 个水平，因子 B 有 s 个水平，现对因子 A 与 B 的不同水平的每种组合下进行试验或抽样，共有 $r\times s$ 个处理，得数据结构见表 5.6.

表 5.6　　无交互作用的双因子方差分析数据结构表

		因子 B				$y_{i\cdot}$
		B_1	B_2	…	B_s	
因子 A	A_1	y_{11}	y_{12}	…	y_{1s}	$y_{1\cdot}$
	A_2	y_{21}	y_{22}	…	y_{2s}	$y_{2\cdot}$
	⋮	⋮	⋮		⋮	⋮
	A_r	y_{r1}	y_{r2}	…	y_{rs}	$y_{r\cdot}$
$y_{\cdot j}$		$y_{\cdot 1}$	$y_{\cdot 2}$	…	$y_{\cdot s}$	

假设 y_{ij} 相互独立，且 $y_{ij}\sim N(\mu_{ij},\ \sigma^2)$，则

$$y_{ij}=\mu_{ij}+\epsilon_{ij}\quad (i=1,2,\cdots,r;\ j=1,2,\cdots,s),$$

其中 ε_{ij} 独立同分布，且 $\varepsilon_{ij} \sim N(0,\sigma^2)$，记

$$\bar{\mu} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij},\quad \bar{\mu}_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s}\mu_{ij},\quad \alpha_i = \bar{\mu}_{i\cdot} - \bar{\mu},\quad \bar{\mu}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r}\mu_{ij},\quad \beta_j = \bar{\mu}_{\cdot j} - \bar{\mu},$$

称 $\bar{\mu}$ 为**总平均值**，称 α_i 为**因素 A 在水平 i 下对实验指标的效应值**，β_j 为**因素 B 在水平 j 下对实验指标的效应值**，显然有 $\sum_{i=1}^{r}\alpha_i = 0$，$\sum_{j=1}^{s}\beta_j = 0$. 于是，可概括双因子方差分析数学模型如下.

假定

$$\begin{cases} y_{ij} = \bar{\mu} + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i(j) = 1,2,\cdots,r(s)), \\ \sum_{i=1}^{r}\alpha_i = 0, \quad \sum_{j=1}^{s}\beta_j = 0, \\ \varepsilon_{ij} \sim N(0,\sigma^2) \text{ 且相互独立}, \end{cases}$$

系统分析因子 A 和因子 B 对实验指标影响的大小，即在给定的显著性水平 α 下，检验如下统计假设：

$H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ （即因子 A 对实验指标影响不显著）；

$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_s = 0$ （即因子 B 对实验指标影响不显著）.

欲检验假设 H_{01} 或 H_{02}，其检验方法类似于单因子方差分析，利用平方和分解中的各种离差平方和，构造 F 统计量. 记

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{s}(y_{ij} - \bar{y})^2,\quad \bar{y} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}y_{ij};$$

$$SSA = s\sum_{i=1}^{r}(\bar{y}_{i\cdot} - \bar{y})^2,\quad SSB = r\sum_{j=1}^{s}(\bar{y}_{\cdot j} - \bar{y})^2;\quad \bar{y}_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s}y_{ij},\quad \bar{y}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r}y_{ij};$$

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{s}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2.$$

称 SST 为**总偏差平方和**；SSE 为**误差平方和**；SSA，SSB 分别为**因子 A，B 的偏差平方和**. 样本总数 $n = rs$.

同样不加证明地得到下面的两个定理.

定理 5.6（平方和分解定理 2） 在无交互作用的两个因素的方差分析模型中，有

$$SST = SSA + SSB + SSE.$$

定理 5.7（各类平方和分布定理） 在无交互作用的两个因素的方差分析模型中，有

$$\frac{SST}{\sigma^2} \sim \chi^2(n-1),$$

$$\frac{SSA}{\sigma^2} \sim \chi^2(r-1),\quad \frac{SSB}{\sigma^2} \sim \chi^2(s-1),$$

$$\frac{SSE}{\sigma^2} \sim \chi^2((r-1)(s-1)),$$

其中，σ^2 为模型方差.

这两个定理的证明参见文献[2].

通常把以上求统计量的计算列成一张表格，便于结果分析(见表 5.7).

表 5.7　　无交互作用的双因子方差分析表

方差来源	平方和	自由度	均方差	F 值
因子 A	SSA	$f_A = r-1$	$\frac{SSA}{f_A}$	$F_A = \frac{SSA}{f_A} \Big/ \frac{SSE}{f_E}$
因子 B	SSB	$f_B = s-1$	$\frac{SSB}{f_B}$	$F_B = \frac{SSB}{f_B} \Big/ \frac{SSE}{f_E}$
误差 E	SSE	$f_E = (r-1)(s-1)$	$\frac{SSE}{f_E}$	
总和 T	SST	$f_T = rs-1$		

根据 F 值推断 H_{01}，H_{02} 正确与否，决策准则是：当 $F_A > F_\alpha(f_A, f_E)$ 时，则拒绝 H_{01}，否则接受 H_{01}；当 $F_B > F_\alpha(f_B, f_E)$ 时，则拒绝 H_{02}，否则接受 H_{02}. 或者由检验的最小显著性概率 p 作出决策，当 $p < \alpha$ 时拒绝相应的原假设.

5.3.2　有交互作用的双因子方差分析

在许多情况下，两因素之间存在着一定程度的交互作用. 所谓交互作用，就是因素之间的联合搭配作用对实验结果产生了影响. 例如有些合金，当单独加入元素 A 或元素 B 时，性能变化不大，但当两者同时加入，合金性能的变化就特别显著. 在多因素的方差分析中，把交互作用当成一个新因素来处理. 为了考查因素间的交互作用，要求两个方面因素的每一交叉项要有重复实验. 如例 5.6 中，对于不同的树种和地区，每一交叉项都有 5 个试验观测数据. 一般地，在有重复实验的双因于方差分析的这种情况下，数据结构见表 5.8.

表 5.8　　有交互作用的双因子方差分析数据结构表

		因子 B												
		B_1				B_2				…	B_s			
因子 A	A_1	y_{111}	y_{112}	…	y_{11n}	y_{121}	y_{122}	…	y_{12n}	…	y_{1s1}	y_{1s2}	…	y_{1sn}
	A_2	y_{211}	y_{212}	…	y_{21n}	y_{221}	y_{222}	…	y_{22n}	…	y_{2s1}	y_{2s2}	…	y_{2sn}
	⋮	⋮	⋮		⋮	⋮	⋮		⋮		⋮	⋮		⋮
	A_r	y_{r11}	y_{r12}	…	y_{r1n}	y_{r21}	y_{r22}	…	y_{r2n}	…	y_{rs1}	y_{rs2}	…	y_{rsn}

表中数据 y_{ijk}表示因子A,B 在第i,j 个水平状态下第k 个样本观测值.

与无交互作用的情形比较，有交互作用的双因子方差分析模型一个关键性的变化，就是在考虑各因子效应的同时，还要考虑因子间的交互效应，通常用 $A\times B$ 表示因子间的交互作用.

下面的讨论中，μ_{ij}，$\bar{\mu}$，α_i，β_j 的意义同 5.3.1 节，记 $\gamma_{ij}=(\mu_{ij}-\bar{\mu})-\alpha_i-\beta_j$，$\mu_{ij}-\bar{\mu}$ 反映水平组合(A_i，B_j)对实验指标的总效应，γ_{ij}等于总效应减去A_i 的效应α_i 及B_j 的效应 β_j，所以 γ_{ij}表示A_i 与B_j 对实验指标的交互效应. 于是，有交互作用的双因子方差分析模型如下.

假定

$$\begin{cases} y_{ijk}=\bar{\mu}+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}, \\ \sum\limits_{i=1}^{r}\alpha_i=0, \quad \sum\limits_{j=1}^{s}\beta_j=0, \quad \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s}\gamma_{ij}=0, \\ \varepsilon_{ijk}\sim N(0,\sigma^2)\text{ 且相互独立}, \end{cases}$$

其中，$i=1,2,\cdots,r$；$j=1,2,\cdots,s$；$k=1,2,\cdots,n$. 系统分析因子 A，B 及交互作用对实验指标影响的大小，即在给定的显著性水平 α 下，检验如下统计假设：

$H_{01}:\alpha_1=\alpha_2=\cdots=\alpha_r=0$ （即因子 A 对实验指标影响不显著）；

$H_{02}:\beta_1=\beta_2=\cdots=\beta_s=0$ （即因子 B 对实验指标影响不显著）；

$H_{03}:\gamma_{ij}=0\ (i=1,2,\cdots,r;\ j=1,2,\cdots,s)$ （即 $A\times B$ 对实验指标影响不显著）.

类似于无交互效应的方差分析讨论，其理论公式和推导不再赘述，详见文献[8]，[9]，[10]，[11]. 这里仅列出方差分析表，见表 5.9.

表 5.9　　有交互作用的双因子方差分析表

方差来源	平方和	自由度	均方差	F 值
因子 A	SSA	$f_A=r-1$	$\frac{SSA}{f_A}$	$F_A=\frac{SSA}{f_A}\Big/\frac{SSE}{f_E}$
因子 B	SSB	$f_B=s-1$	$\frac{SSB}{f_B}$	$F_B=\frac{SSB}{f_B}\Big/\frac{SSE}{f_E}$
交互效应 $A\times B$	$SSAB$	$f_{AB}=(r-1)(s-1)$	$\frac{SSAB}{f_{AB}}$	$F_{AB}=\frac{SSAB}{f_{AB}}\Big/\frac{SSE}{f_E}$
误差 E	SSE	$f_E=rs(n-1)$	$\frac{SSE}{f_E}$	
总和 T	SST	$f_T=rsn-1$		

其中

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n}(y_{ijk}-\bar{y})^2,\quad SSA=sn\sum_{i=1}^{r}(y_{i\cdot\cdot}-\bar{y})^2,\quad SSB=rn\sum_{j=1}^{s}(\bar{y}_{\cdot j\cdot}-\bar{y})^2,$$

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij\cdot})^2,\quad SSAB=SST-SSA-SSB-SSE,$$

$$\bar{y}=\frac{1}{rsn}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n}y_{ijk},\quad \bar{y}_{i\cdot\cdot}=\frac{1}{sn}\sum_{j=1}^{s}\sum_{k=1}^{n}y_{ijk},\quad \bar{y}_{\cdot j\cdot}=\frac{1}{rn}\sum_{i=1}^{r}\sum_{k=1}^{n}y_{ijk},\quad \bar{y}_{ij\cdot}=\frac{1}{n}\sum_{k=1}^{n}y_{ijk}.$$

检验准则是：当计算出的 F 值大于给定 α 的临界值时，则拒绝相应的原假设．或者由检验的最小显著性概率 p 作出决策，当 $p<\alpha$ 时拒绝相应的原假设．

双因素方差分析的计算量比较大，我们用数学软件进行计算．

对于双因素方差分析的问题的处理，MATLAB 提供了命令 anova2，其调用格式为

```
[p, table] = anova2(X, reps, 'displayopt')
```

这个命令和 anova1()类似，只是输入矩阵 X 的行、列各表示一个因子，不同的行(列)表示该因子不同处理下的响应变量的观测值向量．每一个“行与列的偶对”称为一个数据单元，如果各数据单元拥有多于一个的观测点，则参数 reps 声明每一个单元观测点的数目．如在下面的矩阵中

$$\begin{array}{cc} A=1 & A=2 \end{array}$$

$$\left.\begin{bmatrix} x_{111} & x_{112} \\ x_{121} & x_{122} \\ x_{211} & x_{212} \\ x_{221} & x_{222} \\ x_{311} & x_{312} \\ x_{321} & x_{322} \end{bmatrix}\begin{matrix} \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}B=1 \\ \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}B=2 \\ \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}B=3 \end{matrix}\right.$$

行因子有三种不同处理，列因子有两种不同处理，每个数据单元不同数据标号(变动的下标)个数为 2，则 reps = 2(亦即每个数据单元行数与列数的较大者)．

输出参数 p 是检验列、行及其交互作用均值相等的最小显著性概率(向量)．

下面回到例 5.6，这是一个双因子问题．树种和地区作为本题的两个因子，对松树的直径都有可能产生影响，并且二者之间还有可能产生交互作用．即有可能出现某个地区最适合(不适合)某种松树的生长．地区因子有 4 个水平，树种因子有 3 个水平，在每一个水平下分别抽取了 5 个样本．我们先利用 MATLAB 提供的命令 anova2()来对本题作双因子方差分析，再用单因子方差分析确定其他问题．

MATLAB 数据处理

```
clear
A = [23 15 26 13 21 25 20 21 16 18 21 17 16 24 27 14 11 19 20 24];
B = [28 22 25 19 26 30 26 26 20 28 19 24 19 25 29 17 21 18 26 23];
C = [18 10 12 22 13 15 21 22 14 12 23 25 19 13 22 18 12 23 22 19];
```

```
X=[A', B', C'];
```

① 双因子方差分析.

```
reps = 5;
[p, Table] = anova2(X, reps, 'off')
```

上述指令的运行结果是：

```
p =
  0.0004      0.3996      0.4156
Table =
  'Source'       'SS'            'df'   'MS'       'F'      'Prob>F'
  'Columns'      [352.5333]      [2]    [176.2667] [9.1369] [4.3408e-004]
  'Rows'         [58.0500]       [3]    [19.3500]  [1.0030] [0.3996]
  'Interaction'  [119.6000]      [6]    [19.9333]  [1.0333] [0.4156]
  'Error'        [926.0000]      [48]   [19.2917]  []       []
  'Total'        [1.4562e+003]   [59]   []         []       []
```

双因子方差分析结果说明：我们看到返回向量 p 有 3 个元素，分别表示输入矩阵 X 的列、行及交互作用的均值相等的最小显著性概率. 由于 X 的列表示树种方面的因素，行表示地区方面的因素，所以根据这 3 个概率值我们可以知道：树种因素方面的差异显著，地区之间的差异和交互作用的影响不显著. 即没有某种树特别适合在某地区种植.

接下来对树种进一步作单因子方差分析.

② 单因子方差分析.

```
[p, anovatab, stats] = anova1(X, [], 'on')
```

上述指令的运行结果见图 5.2 及：

```
p =
 3.7071e-004
anovatab =
'Source'    'SS'             'df'   'MS'         'F'        'Prob>F'
'Columns'   [352.5333]       [2]    [176.2667]   [9.1036]   [3.7071e-004]
'Error'     [1.1036e+003]    [57]   [ 19.3623]   []         []
'Total'     [1.4562e+003]    [59]   []           []         []
stats =
    gnames:  [3x1 char]
         n:  [20 20 20]
    source:  'anova1'
     means:  [19.5500   23.5500   17.7500]
```

```
df:  57
 s:  4.4003
```

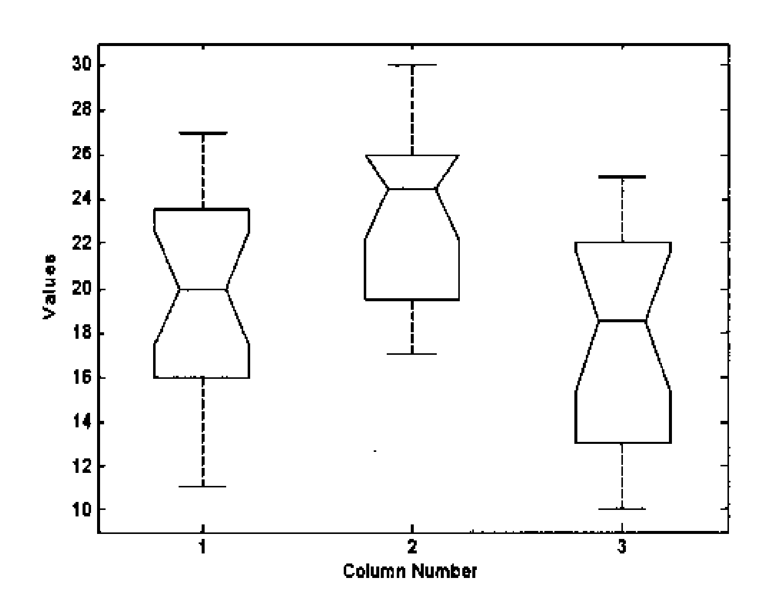


图 5.2　三种松树直径的 box 图

单因子方差分析结果说明：树种 B 的平均直径最大，认为树种 B 最好. 实际上，作多重比较得出的结论更细腻、丰富一些.

习题 5

1. 装配一个部件时可以采用不同的方法，所关心的问题是哪一种方法的效率更高，劳动效率可以用平均装配时间反映. 现从不同的装配方法中各抽取 12 种产品，记录各自的装配时间(单位：min)如表 5.10 所示.

表 5.10

甲方法	31	34	29	32	35	38	34	30	29	32	31	26
乙方法	26	24	28	29	30	29	32	26	31	29	32	28

两个变量为正态分布，且方差相同. 问两种方法的装配时间有无显著不同.($\alpha=0.05$)

2. 为了检验三家工厂生产的机器加工一批原料所需的平均时间是否相同，某化学公司得到了关于加工原料所需时间的数据如表 5.11 所示. 利用这些数据检验三家工厂加工一批原料所需平均时间是否相同.($\alpha=0.05$)

表 5.11

工　厂	1	2	3
加工时间	20	28	20
	26	26	19
	24	31	23
	22	27	22

3. 一项调查研究了信息来源渠道对于信息传播效果的影响. 在该研究中, 信息来源分别为上级、同事和下属. 表 5.12 列出了各种信息渠道的传播效果: 数值越高表示信息传播效果越好. 请检验信息来源对信息传播效果是否有显著影响. ($\alpha=0.05$)

表 5.12

上级	8	5	4	6	6	7	5	5
同事	6	6	7	5	3	4	7	6
下属	6	5	7	4	3	5	7	5

4. 某杂志的一个研究得出这样的结论, 自由职业者的工作压力比非自由职业者的工作压力大. 在该研究中, 为度量一些模糊的概念, 专门设计了若干问题. 这些问题按照从强烈同意到强烈反对分成 1～5 级进行评分, 得分越高表明工作压力越大. 现随机选取三类职业的从业人员: 房地产代理商、建筑师和股票经济人各 15 人, 研究某工作压力, 得到分值如表 5.13 所示.

表 5.13

房地产代理商	建筑师	股票经纪人
81	43	65
48	63	48
68	60	57
69	52	91
54	54	70
62	77	67
76	68	83
56	57	75
61	61	53
65	80	71
64	50	54
69	37	72
83	73	65
85	84	58
75	58	58

对于 $\alpha=0.05$, 检验三种职业的工作压力是否有显著差异.

5. 有 8 位食品专家对三种配方的食品随机品尝, 然后给食品的口感分别打分(满分为 10 分), 见表 5.14. 问三种配方的平均分数是否相同? ($\alpha=0.05$)(假定打分服从标准差相等的正态分布)

表 5.14

专家	1	2	3	4	5	6	7	8
配方 1	8	4	5	6	7	8	6	5
配方 2	6	2	7	5	3	7	4	6
配方 3	5	7	6	3	4	7	5	5

6. 对一所大学的研究生按专业分组，试在显著性水平 $\alpha=0.10$ 下检验他们某科学习成绩是否有明显差异? 表 5.15 列出了考试成绩，假定学生成绩服从方差相等的正态分布.

表 5.15

专　业	成　绩				
甲	75	62	71	58	73
乙	81	85	68	92	90
丙	83	79	60	75	81

7. 某工厂实行早、中、晚三班工作制. 工厂管理部门想了解不同班次工人劳动效率是否存在明显的差异. 每个班次随机抽出了 7 个工人，得工人的劳动效率(单位:件/班)资料见表 5.16. 分析不同班次工人的劳动效率是否有显著性差异.($\alpha=0.05, 0.01$)

表 5.16

早　班	中　班	晚　班
34	49	39
37	47	40
35	51	42
33	48	39
33	50	41
35	51	42
36	51	40

8. 比较 3 种化肥(两种新型化肥 A, B 和传统化肥 C)施撒在三种类型(酸性、中性和碱性)的土地上对作物的产量情况有无差别，将每块土地分成 3 块小区，施用 A, B 两种新型化肥和传统化肥. 收割后，测量各组作物的产量，得到的数据见表 5.17.

表 5.17

化肥种类	酸性土地	中性土地	碱性土地
A	30	31	32
B	31	36	32
C	27	29	28

假定化肥类型与土地类别之间不存在交互效应($\alpha=0.05$). 问:

(1) 化肥对作物产量有影响吗?

(2) 土地类型对作物产量有影响吗?

9. 有三个工人分别在4台机器上加工某种零件, 工作的3天中日产量见表5.18.

表5.18

机器 \ 工人	B_1			B_2			B_1		
A_1	15	15	17	19	19	16	16	18	21
A_2	17	17	17	15	15	15	19	22	22
A_3	15	17	16	18	17	16	18	18	18
A_4	18	20	22	15	16	17	17	17	17

试在显著性水平$\alpha=0.05$下检验操作工人之间技术水平的差异是否显著? 机器性能之间的差异是否显著? 交互作用的影响是否显著?

10. 在某橡胶配方中, 考虑3种不同的促进剂和4种不同分量的氧化剂, 用同样的配方试验两次, 测得300%的定伸强力见表5.19. 试问氧化剂、促进剂以及它们的交互作用对定伸强力有无显著影响? ($\alpha=0.05$)

表5.19

促进剂	氧化剂							
	B_1		B_2		B_3		B_4	
A_1	31	33	34	36	35	36	39	38
A_2	33	34	36	37	37	39	38	41
A_3	35	37	37	38	39	40	42	44

11. 某SARS研究所对31人进行某项生理指标测试, 结果见表5.20.

表5.20

SARS患者	1.8	1.4	1.5	2.1	1.9	1.7	1.8	1.9	1.8	1.8	2.0
疑似者	2.3	2.1	2.1	2.1	2.6	2.5	2.3	2.4	2.4		
非患者	2.9	3.2	2.7	2.8	2.7	3.0	3.4	3.0	3.4	3.3	3.5

问: 这三类人的该项生理指标有差别吗? 如果有差别, 请进行多重比较分析.($\alpha=0.05$)

12. 为培养职业技术教育的师资, 通过统计分析, 认为招收在职生比招收应届生好. 以往招生只确定一个录取分数线, 对年龄和工龄并没有严格的限制, 形成学生间在生活习惯和兴趣爱好等方面有较大的差异. 对年龄、工龄两因素各取两个水平, 重复作四次交叉试验, 考虑两因素与学习成绩的关系. 年龄、工龄两因素各取两个水平如下:

A_1：年龄不超过 25 岁，A_2：年龄超过 25 岁；

B_1：工龄不到 5 年，B_2：工龄超过 5 年.

对某年级在职生两年来所有课程的平均成绩整理见表 5.21.

表 5.21

	B_1					B_2				
A_1	86	87	76	79	85	82	93	82	88	91
A_2	77	82	84	90	76	82	82	80	75	79

试问年龄、工龄以及它们的交互作用对成绩有无显著影响？($\alpha=0.05$)

第6章 回归分析

在一些实际问题中，经常需要我们从定量的角度去研究某些变量间的关系.

一般来讲，变量间的关系有两类. 一类是函数关系，即变量之间确实存在的且在数量上表现为确定性的相互依存关系. 例如，圆的面积 S 与半径 r 有关，一旦半径 r 确定，则面积 S 可通过函数 $S=\pi r^2$ 求出. 另一类是相关关系，即变量之间确实存在的但在数量上表现为不确定的相互依存关系. 例如，人的体重与身高有关，一般而言，较高的人体重较重，但同样身高的人体重却不会完全相同；又如，居民的储蓄存款额与他的收入有关，但同样收入的人储蓄存款额也不会相同.

在很多情况下，函数关系往往通过具有不确定性的相关关系表现出来，而完全的相关关系必定是函数关系.

回归分析是分析变量间相关关系的一种统计方法. 所谓回归分析，就是建立变量之间相关关系的具体的数学表达形式. 根据相关关系的具体形态，明确谁是自变量(可控变量)、谁是因变量(随机变量)，选择一个合适的数学模型来近似地表达变量间的平均变化关系，并借此来探讨对变量的控制与预测问题. 这不仅依赖对变量之间相关程度的度量(需要相关分析的辅助)，更依赖变量之间真实相关性的存在. 然而，现象之间是否存在真实相关，必须根据有关专业领域的学科理论来确定. 因此，回归分析必须要在定性分析的前提下进行，不能进行纯数量的计算.

本章讨论线性回归分析的基本方法.

6.1 一元线性回归分析

6.1.1 一元线性回归模型

在一元线性回归分析中，通常考虑两个变量：一个是自变量 x，其值是可以控制或精确测量的，认为它是非随机变量；另一个是因变量 y，对给定的 x 值，y 的取值事先不能确定，故 y 是随机变量. 为了研究 y 与 x 之间的相关关系，首先就要对变量偶对 (x, y) 进行观测，收集数据. 为使下面的讨论直观，先来考查一个例子.

【例6.1】 我们知道，营业税税收总额 y 与社会商品零售总额 x 有关. 为能从社会商品零售总额去预测税收总额，需要了解两者的关系. 现收集了如下九组数据，见表6.1.

表 6.1　　社会商品零售总额与税收总额　　单位：亿元

序　号	社会商品零售总额 x	营业税税收总额 y
1	142.08	3.93
2	177.30	5.96
3	204.68	7.85
4	242.88	9.82
5	316.24	12.50
6	341.99	15.55
7	332.69	15.79
8	389.29	16.39
9	453.40	18.45

试分析税收总额 y 与商品零售总额 x 的相关关系，建立回归方程.

通常将上述数据记为$(x_i,\ y_i)$ $(i=1,2,\cdots,n)$，本例 $n=9$. 为了直观起见，可将这 n 对数据作为平面直角坐标系 xOy 中的 n 个点，通过描点在平面上得到一张"散点图"，以观察两个变量之间的线性相关性. 本例的散点图见图 6.1.

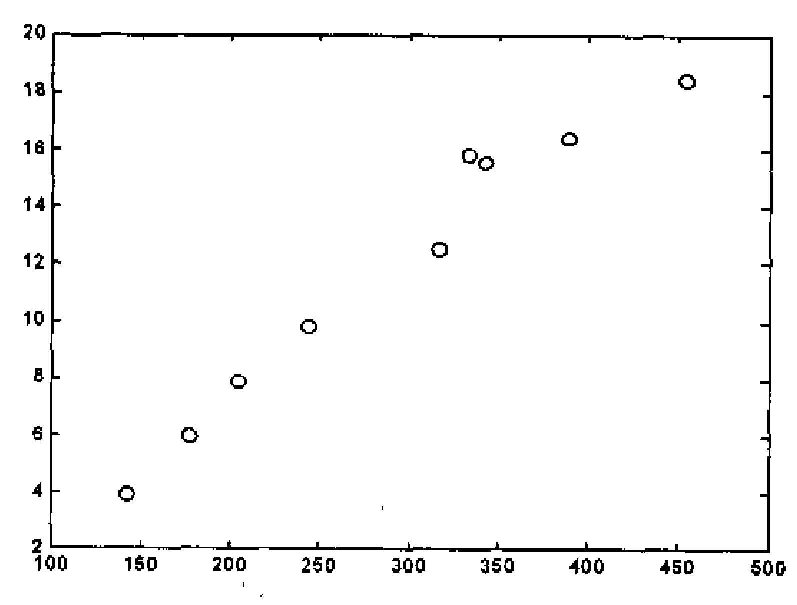


图 6.1　社会商品零售总额与税收总额(x,y)散点图

观察 n 个点在图中的散布情况，发现本例的 9 个点散布在一条直线附近.

我们可以这样理解图中的信息：税收总额 y 与商品零售总额 x 之间似乎存在一种线性关系，也就是说税收总额 y 应当是商品零售总额 x 的线性函数；但是实际观察到的数据点(x_1,y_1)，…，(x_9,y_9)却不在一条直线上，这应当是未知的随机因素干扰的结果. 换句话说，税收总额的观测值 y 由两部分叠加而成：一是税收总额 y 随商品零售总额 x 的变化而呈线性变化的趋势(用 $\alpha+\beta x$ 表示)；另一是其他随机因素干扰的总和(用 ε 表

示)，即观测数据$(x_i,\ y_i)$ $(i=1,2,\cdots,9)$应当满足关系式

$$y_i=\alpha+\beta x_i+\varepsilon_i,$$

通常假定$\varepsilon_i\sim N(0,\sigma^2)$ $(i=1,2,\cdots,9)$且各个ε_i相互独立. 至此，可以给出一元线性回归分析的基本概念.

定义 6.1 (一元线性回归模型)　设x是自变量(非随机变量，其值是可以控制或精确测量的)，y是因变量(随机变量，对给定的x值不能事先确定y的取值)，则称

$$y=\alpha+\beta x+\varepsilon\quad(\varepsilon\sim N(0,\sigma^2))$$

为**一元线性回归模型(理论模型)**. 其中，α,β称为**模型参数**；ε称为**模型随机误差**.

求线性函数

$$E(y)=\alpha+\beta x$$

的经验回归方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x$$

称为**建立一元线性回归模型**. 其中，$\hat{y}$是$E(y)$的统计估计；$\hat{\alpha},\hat{\beta}$分别是α,β的统计估计，称为**经验回归系数**.

设数据对(x_i,y_i) $(i=1,2,\cdots,n)$是变量对$(x,\ y)$的观测数据，则

$$y_i=\alpha+\beta x_i+\varepsilon_i$$

称为**一元样本回归方程(数据模型)**. 其中，$\varepsilon_i\sim N(0,\sigma^2)$ $(i=1,2,\cdots,n)$且各个ε_i相互独立.

6.1.2　模型参数的估计

一元线性回归分析的核心工作是建立一元线性回归模型，其关键是如何利用观测数据估计模型参数，即求出回归系数.

求回归系数的最常用的方法是最小二乘估计. 下面给出最小二乘估计的概念.

定义 6.2 (一元线性最小二乘估计)　称

$$Q(\alpha,\beta)=\sum_{i=1}^{n}[y_i-E(\hat{y}_i)]^2=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

为$y_i(i=1,2,\cdots,n)$回归到直线$E(y)=\alpha+\beta x$时的误差平方和. 若存在$\hat{\alpha}$，$\hat{\beta}$，使得

$$Q(\hat{\alpha},\hat{\beta})=\min_{\alpha,\beta}Q(\alpha,\ \beta),$$

则称$\hat{\alpha}$，$\hat{\beta}$为模型参数α，β的**最小二乘估计**，并称

$$\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i$$

为因变量$y_i(i=1,2,\cdots,n)$的**回归拟合值**，简称**回归值**或**拟合值**. 称

$$e_i=y_i-\hat{y}_i$$

为因变量 $y_i(i=1,2,\cdots,n)$的残差.

关于最小二乘估计的算法，可以从不同的角度推出，应用中最为方便的是矩阵算法.

定理 6.1（一元线性回归模型参数最小二乘估计的矩阵算法） 记

$$\boldsymbol{y}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix}1&x_1\\1&x_2\\\vdots&\vdots\\1&x_n\end{pmatrix},\quad \boldsymbol{A}=\begin{pmatrix}\hat{\alpha}\\\hat{\beta}\end{pmatrix},$$

则一元线性回归的数据模型为 $\boldsymbol{y}=\boldsymbol{XA}$. 这是一个不相容线性方程组，当 $\mathrm{rank}(\boldsymbol{X})=2<n$ 时，其最小二乘解为

$$\boldsymbol{A}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

通常，在高等代数的广义逆矩阵理论中有关于这一算法的详细推证，感兴趣的读者请自行查阅有关教程.

设$\hat{\alpha},\hat{\beta}$为模型参数 α,β 的最小二乘估计，可以证明下面的结论.

① $\hat{\alpha},\hat{\beta}$是 α,β 的无偏估计，即 $E(\hat{\alpha})=\alpha$，$E(\hat{\beta})=\beta$.

② $\hat{\alpha}$和$\hat{\beta}$均服从正态分布，即

$$\hat{\alpha}\sim N\left(\alpha,\ \left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right),\quad \hat{\beta}\sim N\left(\beta,\ \frac{\sigma^2}{l_{xx}}\right),$$

其中

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

由此可知，提高$\hat{\alpha}$和$\hat{\beta}$估计精度的一个基本策略是增加样本容量 n，采样应尽可能分散(即增大 l_{xx}).

③ 参数的区间估计：在结论①和②以及 6.1.3 节关于模型标准差 σ 的估计之上，可推出 α 的 $1-\alpha$ 置信区间为

$$\left(\hat{\alpha}-t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}^*\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}},\ \hat{\alpha}+t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}^*\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}\right),$$

β 的 $1-\alpha$ 置信区间为

$$\left(\hat{\beta}-t_{1-\frac{\alpha}{2}}(n-2)\frac{\hat{\sigma}^*}{\sqrt{l_{xx}}},\ \hat{\beta}+t_{1-\frac{\alpha}{2}}(n-2)\frac{\hat{\sigma}^*}{\sqrt{l_{xx}}}\right),$$

其中，$\hat{\sigma}^*$是 σ 的无偏估计(详见 6.1.3 节).

④ β 的估计值$\hat{\beta}$与 x，y 的相关系数 r_{xy}是成正比例的，即$\hat{\beta}=kr_{xy}(k>0)$. 这一点在应用中对回归方程的解释而言是非常重要的.

相关证明参见文献[2].

6.1.3　回归方程的显著性检验

在回归分析中，一个重要的环节就是分析回归方程的拟合效果(从统计上判断回归方程是否有意义). 这项工作主要的是回归方程的显著性检验. 容易理解，如果变量 y 与变量 x 不存在线性相关关系，此时相关系数 $r_{xy}=0$，即不论 x 如何变化，$E(y)$不会随之而改变，在这种情况下求出的回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$ 是没有意义的. 因此，回归方程显著性检验的问题可描述为 $H_0:\beta=0$. 如果检验不能拒绝 H_0，则得到的回归方程不能用来进一步分析变量 y 与 x 的关系.

回归方程显著性检验的方法很多，这里介绍最常用的方差分析方法.

定义 6.3　在一元线性回归方程的显著性检验中，称统计量

$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 为**总偏差平方和**，其自由度为 $f_T=n-1$；

$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 为**回归平方和**，其自由度为 $f_R=1$；

$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 为**残差平方和**，其自由度为 $f_E=n-2$.

定理 6.2 (偏差平方和分解定理)　$SST=SSR+SSE$.

定理 6.2 的结论是显然的，在多数的数理统计教程中都可以找到其证明.

定理 6.3 (检验统计量构造定理)

① $\dfrac{SSE}{\sigma^2}\sim\chi^2(n-2)$；

② 在 H_0 为真时，$SSR\sim\chi^2(1)$；

③ SSR 与 SSE 相互独立；

④ $F=\dfrac{SSR}{\dfrac{SSE}{n-2}}\sim F(1,n-2)$(检验统计量).

定理 6.3 的结论①、②、③的证明稍难一些，可参见文献[1]；由 F 分布的统计生成定理，结论④是显然的.

于是，根据定理 6.3，回归方程显著性检验的方差分析(F 检验)方法如下.

① 求出回归平方和 SSR 与残差平方和 SSE，进而求出检验统计量 F 的值.

② 求出检验的显著性概率 $p=P\{F(1, n-2)>F\}$.

③ 检验决策，决策准则是：

在显著性水平 α 下，当 $\alpha>p$ 时拒绝 H_0，即认为回归方程有显著意义.

当 $p<0.01$ 时，称回归方程高度显著，标记为 * *；

当 $0.01 \leqslant p < 0.05$ 时，称回归方程显著，标记为 * ；

当 $p \geqslant 0.05$ 时，称回归方程不显著，不作标记.

通常，将检验结果整理成如表 6.2 所示的检验报告(方差分析表).

表 6.2　回归方程显著性检验的方差分析

方差来源	偏差平方和	自由度	F 值	p 值	显著性
回归 残差	SSR SSE	$f_R=1$ $f_E=n-2$	$F=\dfrac{SSR/f_R}{SSE/f_E}$	$p=P\{F(1,n-2)>F\}$	
总计	SST	$f_T=n-1$			

关于回归方程拟合效果的分析，还可以从另外两个方面进行.

① 可决系数分析. 最常用的测定回归直线对各个观测点的拟合程度的统计量是 $r^2=\dfrac{SSR}{SST}$，通常称之为**可决系数**. 显然，$r^2 \in [0,1]$，r^2 的值越大(小)，表明回归直线对各个观测点的拟合程度越高(低). 若 $r^2=1$，即 $SSE=0$，表明 y 对 x 几乎有确定的线性函数关系；若 $r^2=0$，即 $SSR=0$，表明 y 对 x 完全没有线性相关关系. 注意，简单的推导即可明了 $r=\pm\sqrt{r^2}$ 的统计意义，r 等于变量 y 的观测数据 $y_1, y_2, \cdots, y_n$ 与模型拟合数据 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 之间的相关系数，其正负号与回归系数 $\hat{\beta}$ 的正负号相同.

② 估计的标准误差. 由定理 6.3 的结论①可知，$E(SSE)=(n-2)\sigma^2$. 因此定义统计量 $\hat{\sigma}^{*2}=\dfrac{SSE}{n-2}$，显然，$\hat{\sigma}^{*2}$ 是模型方差 σ^2 的无偏估计，进而 $\hat{\sigma}^*=\sqrt{\dfrac{SSE}{n-2}}$ 可以作为对模型标准差 σ 的估计，通常称为变量 y 对 x 的最小二乘回归的**估计标准误差**. 显然，$\hat{\sigma}^*$ 的值越小，表明回归直线对各个观测点的拟合程度越高.

需要指出的是，可以证明 $\hat{\sigma}^2_{\mathrm{MLE}}=\dfrac{1}{n}SSE$ 是 σ^2 的有偏估计，$\hat{\sigma}^{*2}=\dfrac{n}{n-2}\hat{\sigma}^2_{\mathrm{MLE}}$. 由定理 6.3 的结论①，容易得到模型方差 σ^2 的 $1-\alpha$ 置信区间为 $\left(\dfrac{SSE}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}, \dfrac{SSE}{\chi^2_{\frac{\alpha}{2}}(n-2)}\right)$.

6.1.4　利用回归方程进行预测

建立回归方程的目的不仅是描述变量之间的关系，更重要的是回归方程的应用. 利用所建回归方程对因变量进行预测是其应用的基本内容. 在一元线性回归分析中，当回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$ 具有统计显著性时，利用回归方程容易实现对因变量 y 的预测，而这一问题的实质是对 y 的点估计和区间估计.

在 6.1.2 节讨论的基础上，容易证明：

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \sim N\left(\alpha + \beta x, \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{l_{xx}}\right)\sigma^2\right), \text{且} \hat{y} \text{与} y \text{独立}.$$

这个结论表明，经验回归方程$\hat{y} = \hat{\alpha} + \hat{\beta}x$是线性函数$E(y) = \alpha + \beta x$的无偏估计.

因此，当$x = x_0$时，因变量y的预测值即为$\hat{y}_0 = a + bx_0$，它是$y_0 = a + bx_0 + \varepsilon_0$的无偏估计. 在显著性水平$\alpha$下，$y_0$的估计边际误差(区间估计)可由准则式

$$P\{|y_0 - \hat{y}_0| < \delta\} \geqslant 1 - \alpha$$

确定. 由y和$\hat{y}$的分布可以推出

$$\delta = t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$

显然，预测的精度取决于δ的大小，而影响δ大小的因素主要是样本容量n，x_0与x的距离以及自变量的偏差平方和l_{xx}. 当样本容量n较大，x_0与x的距离较近，自变量的偏差平方和l_{xx}较大(采样较为分散)时，δ的取值就较小，此时预测的精度较高. 另外，当$x_0 \notin [x_{(1)}, x_{(n)}]$时，预测精度可能变得很差，在这种情况下作外推，需要特别小心.

由于上面计算边际误差δ的公式略显冗繁，故在实际应用中，当x_0取在$\bar{x}$附近，n很大时，利用$\hat{y}_0 - y_0 \overset{\text{近似}}{\sim} N(0, \hat{\sigma}^{*2})$计算近似的边际误差$\delta^*$，此时$y_0$的0.95预测置信区间近似为$(\hat{y}_0 - 2\delta^*, \hat{y}_0 + 2\delta^*)$，0.99预测置信区间近似为$(\hat{y}_0 - 3\delta^*, \hat{y}_0 + 3\delta^*)$.

MATLAB提供了线性回归模型的建模与评价函数regress. 下面利用这个函数完成例6.1的建模与评价. 首先对函数regress的使用方法进行简单介绍.

函数regress可用于p个自变量、一个因变量的线性回归模型$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon} \sim N(\boldsymbol{O}, \sigma^2\boldsymbol{I})$的建模和模型评价. 其调用格式为

```
[b,bint,r,rint,stats] = regress(y,X,alpha)
```

其中，输入参数X表示p个自变量的n个观测值的$n \times p$矩阵，y表示因变量的n个观测值的$n \times 1$向量，alpha是显著性水平(可以缺省，此时默认为0.05)；输出参数b返回的是模型系数(向量)β的最小二乘估计值，bint是β的100(1 - alpha)%置信区间，r是模型拟合残差(向量)，rint是模型拟合残差的100(1 - alpha)%置信区间，stats包含可决系数R^2的值、方差分析的F统计量的值、方差分析的显著性概率p的值和模型方差σ^2的估计值，bint、r、rint和stats可以缺省.

下面给出例6.1回归分析的MATLAB数据处理.

```
clear
x = [142.08,177.30,204.68,242.88,316.24,341.99,332.69,389.29,453.40]';
y = [3.93,5.96,7.85,9.82,12.50,15.55,15.79,16.39,18.45]';
X = [ones(length(x),1),x]; % 构造自变量观测值矩阵
```

```
[b,bint,r,rint,states] = regress(y, X); % 线性回归建模与评价
b, states % 显示所关心输出参数
```

上述指令的运行结果是：

```
b =
  -2.2610
   0.0487
states =
   0.9625    179.7711    0.0000    1.1315
```

由此可知，回归方程为 $\hat{y} = -2.2610 + 0.0487x$，回归方程高度显著，可决系数 $R^2 = 0.9625$，模型方差的估计 $\hat{\sigma}^2 = 1.1315$.

【例 6.2】 利用例 6.1 关于营业税税收额 y 与商品零售额 x 的回归方程，预测当商品零售额 $x = 300$ 亿元时，营业税税收额 y 为多少亿元.

分析 进行点预测和区间预测. 由于 $x = 300$ 亿元接近商品零售额的平均值，故用近似置信区间进行区间预测，显著性水平取 0.05.

MATLAB 数据处理(接例 6.1 进行)

```
x0 = 300;
y0 = b(1) + b(2) * x0 % 点预测
SSE = sum((y - (b(1) + b(2) * x)).^2); % 计算残差平方和
STD = sqrt(SSE/(length(x) - 2)); % 计算标准误差
DELTA = 2 * STD; % 计算 0.05 显著性水平下的边际误差
ci = [y0 - DELTA, y0 + DELTA] % 0.95 置信区间
```

上述指令的运行结果是：

```
y0 =
  12.3423
ci =
  10.2149    14.4698
```

即当社会商品零售总额为 300 亿元时，营业税平均税收总额的预测值约为 12.3423 亿元，其 0.95 置信区间为(10.2149, 14.4698).

6.1.5 目标函数可线性化的曲线回归分析

在一些实际问题中，变量间的关系并不都是线性的，这时就应该用曲线去进行拟合. 首先要解决的问题就是回归方程中的参数如何估计. 解决这一问题的基本思路是：对于曲线回归建模的非线性目标函数 $y = f(x)$，通过某种数学变换 $\begin{cases} v = v(y), \\ u = u(x) \end{cases}$ 使之"线

性化”，化为一元线性函数 $v = a + bu$ 的形式，继而利用线性最小二乘估计的方法估计出参数 a 和 b，用一元线性回归方程 $\hat{v} = \hat{a} + \hat{b}u$ 来描述 v 与 u 间的统计规律性，然后再用逆变换 $\begin{cases} y = v^{-1}(v), \\ x = u^{-1}(u) \end{cases}$ 还原为目标函数形式的非线性回归方程. 表 6.3 给出了常用的非线性函数及其线性化的方法.

表 6.3　　常用的非线性函数线性化的方法

名　称	定　义	图　像	线性化方法
倒幂函数	$y = a + b\dfrac{1}{x}$	y O x	令 $v = y$，$u = \dfrac{1}{x}$，则 $v = a + bu$
双曲线函数	$\dfrac{1}{y} = a + \dfrac{b}{x}$	y $b>0$ $b<0$ O x	令 $v = \dfrac{1}{y}$，$u = \dfrac{1}{x}$，则 $v = a + bu$
幂函数	$y = ax^b$	y $b<0$ $0<b<1$ $b>1$ O x	令 $v = \ln y$，$u = \ln x$，则 $v = \ln a + bu$
指数函数	$y = ae^{bx}$	y $b<0$ $b>0$ O x	令 $v = \ln y$，$u = x$，则 $v = \ln a + bu$
倒指数函数	$y = ae^{b/x}$	y $b<0$ $b>0$ O x	令 $v = \ln y$，$u = \dfrac{1}{x}$，则 $v = \ln a + bu$

续表 6.3

名　称	定　义	图　像	线性化方法
对数函数	$v=a+b\ln x$	y, b>0, b<0, O, x	令 $v=y$, $u=\ln x$, 则 $v=a+bu$
S 型曲线	$y=\dfrac{1}{a+be^{-x}}$	y, O, x	令 $v=\dfrac{1}{y}$, $u=e^{-x}$, 则 $v=a+bu$

当目标函数线性化之后，接下来的线性回归建模同前，模型评价工作应在线性回归方程还原为非线性回归方程后进行，相关概念和公式同线性回归，这里不再赘述.

【例 6.3】 为了解百货商店销售额 x 与流通费率(反映商业活动的一个质量指标，指每元商品流转额所分摊的流通费用) y 之间的关系，收集了九个商店的有关数据，见表 6.4. 试建立流通费率 y 关于销售额 x 的回归方程.

表 6.4　　销售额与流通费率数据

样本点	销售额 x/万元	流通费率 y/%
1	1.5	7.0
2	4.5	4.8
3	7.5	3.6
4	10.5	3.1
5	13.5	2.7
6	16.5	2.5
7	19.5	2.4
8	22.5	2.3
9	25.5	2.2

分析　首先绘制散点图以直观地选择拟合曲线，这项工作应结合相关专业领域的知识和经验进行，有时可能需要多种尝试. 选定目标函数后进行线性化变换，针对变换后的线性目标函数进行回归建模与评价，然后再还原为非线性回归方程.

MATLAB 数据处理

① 绘制散点图以直观地选择拟合曲线.

```
clear
```

```
x=[1.5,4.5,7.5,10.5,13.5,16.5,19.5,22.5,25.5];
y=[7.0,4.8,3.6,3.1,2.7,2.5,2.4,2.3,2.2];
plot(x,y,'-o')
```

上述指令的运行结果见图6.2.

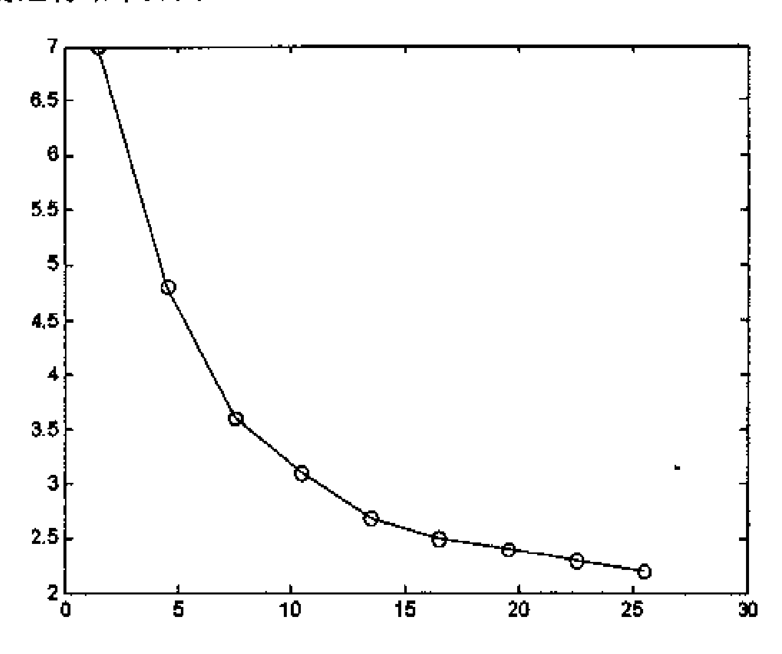


图6.2 销售额与流通费率数据散点图

对比图6.2与表6.3中所列函数图像，初步判断应以幂函数曲线为拟合目标，即非线性回归建模的目标函数为$y=ax^b(b<0)$，其线性化变换公式为$v=\ln y$，$u=\ln x$，线性函数为$v=\ln a+bu$.

② 线性化变换即线性回归建模与模型评价.

```
U=log(x)'; % 线性化变换
V=log(y)'; % 线性化变换
MU=[ones(length(U),1),U]; % 构造自变量观测值矩阵
[b,bint,r,rint,states]=regress(V,MU); % 线性回归建模评价
b, states
```

上述指令的运行结果是：

```
b =
    2.1421
   -0.4259
states =
    0.9928    963.5572    0.0000    0.0012
```

由此可知，回归方程为$\hat{y}=2.1421-0.4259x$，回归方程高度显著，可决系数$R^2=$

0.9928，模型方差的估计$\hat{\sigma}^2=0.0012$.

严格来讲，模型评价工作应在逆线性化变换后进行．但是，若所建线性回归方程不理想，则相应的非线性回归方程必定不理想．现得到的线性回归方程是非常理想的．

③ 逆线性化变换求非线性回归方程．

```
A = exp(b(1))%逆线性化变换，模型系数还原
B = b(2)
```

上述指令的运行结果是：

```
A =
      8.5173
B =
     -0.4259
```

即非线性回归方程是$\hat{y}=8.5173x^{-0.4259}$．非线性回归方程的评价略．

6.2　多元线性回归分析

6.2.1　多元线性回归模型

多元线性回归分析是应用最广泛的多元分析方法之一．多元线性回归分析的原理与一元线性回归分析完全相同，但在计算上要复杂得多，通常需要借助计算机和统计软件才能得以应用．

定义 6.4（多元线性回归模型）　设 $x_1, x_2, \cdots, x_p$ 是$p(\geqslant 2)$个自变量(解释变量)，y 是因变量，则称

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon\quad(\varepsilon\sim N(0,\sigma^2))$$

为**多元线性回归模型(理论模型)**．其中，$\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 是 $p+1$ 个**模型参数**(β_0 称为**常数项**，$\beta_1,\beta_2,\cdots,\beta_p$ 称为**模型系数**)；$\varepsilon\sim N(0,\ \sigma^2)$是**模型随机误差**．

求 p 元线性函数

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

的经验回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$$

称为**建立多元线性回归模型**．其中，$\hat{y}$是 $E(y)$的统计估计；$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$ 分别是β_0，$\beta_1,\beta_2,\cdots,\beta_p$ 的统计估计，称为**经验回归系数**．

设对变量向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$，$\boldsymbol{y}$ 的 n 次观测得到的样本数据为$(x_{i1},x_{i2},\cdots,x_{ip},\ y_i)$

$(i=1,2,\cdots,n;\ n>p+1)$. 为了今后讨论方便，引入矩阵

$$y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad X=\begin{pmatrix}1&x_{11}&\cdots&x_{1p}\\1&x_{21}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\1&x_{n1}&\cdots&x_{np}\end{pmatrix},\quad \hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\\hat{\beta}_1\\\vdots\\\hat{\beta}_p\end{pmatrix},\quad \varepsilon=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix},$$

于是，称

$$y=X\beta+\varepsilon$$

为**多元样本回归方程(数据模型)**. 其中，$\mathrm{rank}(X)=p+1<n$，$\varepsilon\sim N_n(O_{n\times1},\ \sigma^2 I_{n\times n})$ 且各个 ε_i 相互独立. 由于矩阵 X 是样本数据，X 的数据可以进行设计和控制，因此，矩阵 X 称为**回归设计矩阵**或**资料矩阵**.

对于多元线性回归模型，需要强调指出的是：

① 条件 $\mathrm{rank}(X)=p+1<n$ 表明，X 是一个满秩矩阵，即矩阵 X 的列向量(解释变量)间线性无关，样本容量的个数应当大于解释变量的个数. 违反该假设时，称**模型存在多重共线性问题**.

② 条件 $\varepsilon\sim N_n(O_{n\times1},\ \sigma^2 I_{n\times n})$且各个 ε_i 相互独立表明，系统受到零均值齐性方差的正态随机干扰，系统自变量之间不存在序列相关，即

$$E(\varepsilon_i)=0,\quad \mathrm{Cov}(\varepsilon_i,\ \varepsilon_j)=\begin{cases}\sigma^2, & i=j,\\0, & i\neq j.\end{cases}\quad (i,j=1,2,\cdots,n)$$

当 $\mathrm{Var}(\varepsilon_i)\neq\mathrm{Var}(\varepsilon_j)$ $(i\neq j)$时，称**回归模型存在异方差**. 当 $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)\neq0$ $(i\neq j)$时，称**回归模型存在自相关**.

当模型违反上述假设后，就不能使用最小二乘法估计回归系数. 解决方法将在后面介绍，先介绍模型符合假设时的参数估计方法.

6.2.2　模型参数的估计

多元线性回归分析估计模型参数的原理和方法同一元线性回归分析.

定义 6.5 (多元线性最小二乘估计)　称

$$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}[y_i-E(y_i)]^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip})^2$$

为**回归离差平方和**. 若存在$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$，使得

$$Q(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)=\min_{\beta_0,\beta_1,\cdots,\beta_p}Q(\beta_0,\beta_1,\cdots,\beta_p),$$

则称$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$ 为模型参数$\beta_0,\beta_1,\cdots,\beta_p$ 的**最小二乘估计**. 称

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip}$$

为因变量 $y_i(i=1,2,\cdots,n)$的**回归拟合值**，简称**回归值**或**拟合值**. 称

$$e_i=y_i-\hat{y}_i$$

为因变量 $y_i(i=1,2,\cdots,n)$的**残差**.

定理 6.4 (多元线性回归模型参数最小二乘估计的矩阵算法) 当满足多元线性回归模型的理论假定时，模型参数 $\beta_1,\beta_2,\cdots,\beta_p$ 最小二乘估计的矩阵算法是

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

可以证明，模型参数的最小二乘估计服从正态分布，即

$$\hat{\beta}_j\sim N(\beta_j,c_{jj},\sigma^2)\quad(j=1,2,\cdots,p),$$

其中$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}=(c_{ij})_{p\times p}$.

由此可见，$\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)^{\mathrm{T}}$ 是 $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 的无偏估计. 协方差阵 $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$反映出估计量$\hat{\boldsymbol{\beta}}$的波动大小，由于 $\mathrm{Cov}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$，所以$\hat{\boldsymbol{\beta}}$的波动大小可以由抽样过程中进行控制. 同一元线性回归分析一样，在多元线性回归中，样本容量要尽可能大，采样要尽可能分散.

6.2.3 回归方程的显著性检验

多元回归方程的显著性较一元的情形要复杂一些.

6.2.3.1 多元回归方程显著性的整体性检验

检验自变量 $x_1,x_2,\cdots,x_p$ 的全体对因变量 y 是否有显著影响，最常用的整体性检验方法仍是方差分析方法. 检验的原假设是 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$(回归方程无意义).

定义 6.6 在多元线性回归方程的显著性检验中，称统计量

$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 为**总偏差平方和**，其自由度为 $f_T=n-1$;

$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 为**回归平方和**，其自由度为 $f_R=p$;

$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 为**残差平方和**，其自由度为 $f_E=n-p-1$.

可以证明，偏差平方和分解定理仍然成立.

定理 6.5 (偏差平方和分解定理) $SST=SSR+SSE$.

进而，可以证明下面的定理.

定理 6.6 (检验统计量构造定理)

① $\frac{SSE}{\sigma^2} \sim \chi^2(n-p-1)$;

② 在 H_0 为真时，$\frac{SSR}{\sigma^2} \sim \chi^2(p)$;

③ SSR 与 SSE 相互独立;

④ $F = \dfrac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F(p, n-p-1)$ （检验统计量）.

于是，根据定理 6.6，多元线性回归方程显著性检验的方差分析方法如下.

① 求出回归平方和 SSR 与残差平方和 SSE，进而求出检验统计量 F.

② 求出检验的显著性概率 $p = P\{F(p, n-p-1) > F\}$.

③ 检验决策，决策准则是:

在显著性水平 α 下，当 $\alpha > p$ 时拒绝 H_0，即认为回归方程有显著意义.

当 $p<0.01$ 时，称回归方程高度显著，标记为 * *;

当 $0.01 \leqslant p < 0.05$ 时，称回归方程显著，标记为 *;

当 $p \geqslant 0.05$ 时，称回归方程不显著，不作标记.

亦应将检验结果整理成方差分析表，如表 6.2 所示.

此外，与一元线性回归分析类似，可用可决系数 $r^2 = \frac{SSR}{SST}$ 来测定回归方程对各个观测点的拟合程度.

由于 $E(SSE) = (n-p-1)\sigma^2$，所以可用统计量 $\hat{\sigma}^{*2} = \frac{SSE}{n-p-1}$ 对模型方差 σ^2 进行估计.

6.2.3.2 多元线性回归方程中每个自变量对因变量影响显著性检验

在多元线性回归分析中，关于自变量对因变量影响显著性的问题，除前面的整体性检验外，通常还要检验每个自变量 x_j 对因变量 y 影响的显著性. 检验的原假设是

$$H_{0j}: \beta_j = 0 \quad (j = 1, 2, \cdots, p).$$

这里扼要介绍常用的 F 检验方法. 检验统计量构造及其分布结论如下.

在 H_{0j} 为真时，检验统计量

$$F_j = \frac{\hat{\beta}_j^2}{c_{jj}\frac{SSE}{n-p-1}} \sim F(1, n-p-1).$$

检验的显著性概率

$$p = P\{F(1, n-p-1) > F_j\}.$$

检验的决策准则是：在显著性水平 α 下，当 $\alpha > p$ 时拒绝 H_{0j}，即认为解释变量 x_j 对因变量 y 影响显著.

若存在不显著的变量，取 $F_k = \min\limits_{1 \leqslant j \leqslant p} \{F_j\}$，从回归方程中剔除自变量 x_k. 设从原回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{k-1} x_{k-1} + \hat{\beta}_k x_k + \hat{\beta}_{k+1} x_{k+1} + \cdots + \hat{\beta}_p x_p$$

中剔除自变量 x_k 后，重新建立的回归方程为

$$\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^* x_1 + \cdots + \hat{\beta}_{k-1}^* x_{k-1} + \hat{\beta}_{k+1}^* x_{k+1} + \cdots + \hat{\beta}_p^* x_p,$$

则可以证明，新回归方程的系数与原回归方程的系数有如下关系：

$$\hat{\beta}_j^* = \hat{\beta}_j - \frac{c_{kj}}{c_{kk}} \hat{\beta}_k \quad (j = 1, 2, \cdots, p;\ j \neq k),$$

$$\hat{\beta}_0^* = \bar{y} - \sum_{j \neq k} \hat{\beta}_j^* \bar{x}_j.$$

对于新建立的回归方程，必须对每一个余下的变量再次进行检验，直至余下变量全部显著为止.

在问题能够满足模型理论的假定条件时，建模与模型评价的数据处理可由前面介绍过的 regress 函数完成. 但是，这种情况在实际应用中是可遇不可求的. 因此，多元线性回归分析更有效的建模方法将在 6.2.5 和 6.3 节中进行讨论.

6.2.4 利用回归方程进行预测

在多元线性回归分析中，当回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ 具有统计显著性时，利用回归方程容易实现对因变量 y 的预测，其方法同一元的情形，这里仅作扼要介绍.

设预测点为 $\boldsymbol{x}_0 = (x_{01}, x_{02}, \cdots, x_{0p})^{\mathrm{T}}$，则

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p}$$

是对

$$E(y_0) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_p x_{0p}$$

的点估计，亦是对

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_p x_{0p} + \varepsilon_0 \quad (\varepsilon_0 \sim N(0, \sigma^2))$$

的点预测. 并且，可以证明统计量

$$t = \frac{y_0 - \hat{y}_0}{\hat{\sigma}^* \Delta} \sim t(n - p - 1),$$

其中

$$\hat{\sigma}^{*2}=\frac{SSE}{n-p-1},\quad \Delta=\sqrt{1+\frac{1}{n}+\sum_{i=1}^{p}\sum_{j=1}^{p}(x_{0i}-\bar{x}_i)(x_{0j}-\bar{x}_j)c_{ij}},$$

$$\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ki}\quad(i=1,2,\cdots,p).$$

于是，点预测的边际误差为$\pm t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}^*\Delta$，即在$x_0$处的区间预测为

$$(\hat{y}_0-t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}^*\Delta,\hat{y}_0+t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}^*\Delta),$$

即

$$P\left\{\hat{y}_0-t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}^*\Delta<y_0<\hat{y}_0+t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}^*\Delta\right\}\geqslant1-\alpha.$$

当n较大，$x_{0i}\approx\bar{x}_i(i=1,2,\cdots,p)$时，可取$\Delta=1$来简化计算．关于这部分内容的详细讨论参见文献[12]．

6.2.5 最优回归方程的选择

在多元线性回归模型的应用中，模型的假定条件往往不能满足，这一点会体现到回归方程的显著性检验结果中．因此，如何通过自变量的筛选以提高回归方程的显著性以至找到最优回归方程是人们关心的问题．

什么是最优回归方程？这在理论上尚无一个明确的标准．但是，在选择所谓的最优回归方程时，下面几点应予考虑：

① 变量完备，回归方程中尽可能包含对因变量有实际影响的自变量；

② 模型从简，回归方程中所包含的自变量的个数尽可能少；

③ 充分拟合，回归方程的剩余方差尽可能小．

显然，这几点在实践中可能出现“跷跷板”现象．因此，根据统计分析和问题的实际背景求得某种平衡才是最优回归方程概念的实质．单从统计分析的角度，人们常用的选择最优回归方程的方法是逐步回归法．方法的操作要点如下：

① 根据问题所属专业领域的理论和经验提出对因变量可能有影响的所有自变量；

② 计算每一个自变量对因变量的相关系数，按其绝对值从大到小排序；

③ 取相关系数绝对值最大的那个自变量建立一元线性回归模型，检验所得回归方程的显著性，若检验表明回归效果显著则转入④，若检验表明回归效果不显著则停止建模；

④ 进行变量的追加、剔除和回归方程的更新操作．

若检验表明回归效果显著，则按相关系数绝对值由大到小的顺序逐一将相应的自变量引入回归方程；每引入一个新的自变量，对新回归方程中每一个自变量都要进行显著性检验．

若检验表明回归效果不显著，则剔除对因变量影响最小的自变量，更新回归方程；对更新后的回归方程中的每一个自变量仍要进行显著性检验、剔除、更新，直到回归方程中的每一个自变量都显著为止，再引入前面未曾引入的自变量.

依此类推，直到无法剔除已经引入的自变量也无法引入新的自变量为止.

需要指出的是，逐步回归法不能保证得到真正的最优回归方程，但此法是计算量较小、预测效果较好、有工具软件支持、应用最多的一种方法. 另外，逐步回归法受检验的显著性水平 α 影响较大，α 较大将会有较多的自变量引入回归方程，α 较小将会导致一些重要的自变量被剔除.

MATLAB 提供了两个用逐步回归法建立多元线性回归模型的函数 stepwisefit 和 stepwise，这两个函数的功能是一样的. 前者是逐步回归法建模的集成命令，使用者只需给出必要的输入参数，调用这一函数将自动完成建模工作，返回所谓最优回归方程的相关信息；后者是逐步回归法建模的交互式图形环境创建指令.

下面简要介绍 stepwisefit 函数的使用方法，stepwise 函数的使用方法参见附录 B.

stepwisefit 函数完整的调用格式是

```
[b,se,pval,inmodel,stats,nextstep,history] =
      stepwisefit(X,y,'Param1',value1,'Param2',value2,...)
```

其中，输入参数

X 是 p 个自变量的 n 个观测值的 $n\times p$ 矩阵.

y 是因变量的 n 个观测值的 $n\times 1$ 向量.

'Paramk'是第 k 个引用参数，valuek 是其取值，通常可以缺省. 这里只介绍 3 个可能会用到的引用参数：

'penter'设置回归方程显著性检验的显著性概率上限，缺省设置为 0.05；

'premove'设置回归方程显著性检验的显著性概率下限，缺省设置为 0.10；

'display'用来指明是否强制显示建模过程信息，取值为'on'(显示，缺省设置)和'off'(不显示).

输出参数

b 是模型系数.

se 是模型系数的标准误差.

pval 是显著性检验各个自变量的显著性概率.

inmodel 是各个自变量在最终回归方程中地位的说明(1 表示在方程中，0 表示不在方程中).

stats 是一个构架数组，包括：

source：建模方法的说明，'stepwisefit'表示逐步回归法；

dfe：最优回归方程的剩余自由度；

表 6.5　　水泥中的化学成分含量与水泥凝固时的放热量数据

序号	x_1	x_2	x_3	x_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

试用逐步回归法求出 Y 对 x_1, x_2, x_3 和 x_4 的最优回归方程.

此例选自 MATLAB 系统帮助，数据保存在 hald.mat 文件中，ingredients 为自变量，heat 为因变量.

```
clear
load hald
[b, se, pval, inmodel, stats, nextstep, history] = stepwisefit (ingredients,
heat,'penter',0.10,'display','off');
inmodel,b0 = stats.intercept,b % 自变量的筛选和模型参数估计信息
ALLp = stats.pval, rmse = stats.rmse % 回归方程显著性整体检验信息
P = stats.PVAL % 回归方程显著性分别检验信息
```

上述指令的运行结果是：

```
inmodel =
        1     1     0     0
b0 =
        52.5773
b =
        1.4683
        0.6623
        0.2500
       -0.2365
```

```
ALLp =
        4.4066e-009
rmse =
        2.4063
P =
        0.0000
        0.0000
        0.2089
        0.2054
```

结果表明，最优回归方程为$\hat{y}=52.5773+1.4683x_1+0.6623x_2$，回归方程显著性整体检验和分别检验均为高度显著，模型标准误差估计为2.4063.

6.3 偏最小二乘回归分析

经典多元线性回归分析(MLR)是研究变量之间相关关系的基本方法．但是，下面两个问题制约着其应用的效能：一是样本容量要求很高，一般应大于30或大于自变量数的5～10倍；二是消除变量间多重相关性很难．若在变量之间存在严重多重相关性，将对回归建模与模型分析工作带来如下危害．

① 在自变量间存在严重多重相关性的情况下，将造成回归资料矩阵的严重病态性，进而使模型参数的最小二乘估计失真．回归系数的估计方差将随着自变量之间相关程度的不断增强而迅速扩大，回归系数的估计值对样本数据的微小变化变得非常敏感，回归系数估计值的稳定性将变得很差．

② 在自变量高度相关的条件下，用最小二乘法得到的回归模型其回归系数的物理含义很难解释．许多从专业知识上看似乎是十分重要的变量，其回归系数的取值变得微不足道，甚至还会出现回归系数的符号与人们的实际概念完全相反的现象．

③ 存在严重的多重共线性影响时，回归系数的统计检验将难以通过．

回归建模过程中必须要解决多重共线性问题．常见的方法是用逐步回归法来进行变量的筛选，去掉不太重要的相关性变量．然而，逐步回归法存在下列问题：一是缺乏对变量间多重相关性进行判定的十分可靠的检验方法；二是删除部分多重相关变量的做法常导致增大模型的解释误差，将本应保留的系统信息舍弃，使得接受错误结论的可能以及作出错误决策的风险不断增长．

在克服变量多重相关性对系统回归建模干扰的努力中，1983年，瑞典的S.Wold和C.Albano等人提出了**偏最小二乘回归分析**(PLS)方法，它开辟了一种有效的技术途径，

在处理样本容量小、解释变量个数多、变量间存在严重多重相关性问题方面具有独特的优势，并且可以同时实现回归建模、数据结构简化以及两组变量间的相关分析.

6.3.1 偏最小二乘回归方法的数据结构与建模思想

设有 q 个因变量 $y_1, y_2, \cdots, y_q$ 与 p 个自变量 $x_1, x_2, \cdots, x_p$，为了研究因变量与自变量的统计关系，观测了 n 个样本点，由此分别构成了自变量与因变量的"样本点×变量"型的数据矩阵，记为

$$X = (x_{ij})_{n \times p} = (x_1, x_2, \cdots, x_p)$$

和

$$Y = (y_{ij})_{n \times q} = (y_1, y_2, \cdots, y_q).$$

PLS 方法在建模过程中采用了信息综合与筛选技术，不直接考虑因变量系统 Y 对自变量系统 X 的回归建模，而是从自变量系统 X 中逐步提取 m 个对自变量系统 X 和因变量系统 Y 都具有最佳解释能力的新综合变量 $t_1, \cdots, t_m (m \leqslant p)$，亦称之为主成分. 首先建立 y_k 对主成分 $t_1, \cdots, t_m$ 的 MLR 回归方程，然后还原为 y_k 关于原自变量系统 $x_1, x_2, \cdots, x_p$ 的 PLS 回归方程，其中 $k = 1, 2, \cdots, q$.

PLS 方法的关键性技术是提取主成分，基本思想如下.

第一步，分别在 X 和 Y 中提取第一主成分 t_1 和 u_1，并且要求：

① 主成分的代表性，t_1 和 u_1 应尽可能大地携带各自的变量系统中的变异信息；

② 主成分的相关性，t_1 和 u_1 的相关程度能够达到最大，即 t_1 对因变量系统有很强的解释能力.

这两个要求表明，PLS 方法主成分的提取同主成分分析中主成分的提取既有相似之处(代表性要求)，又有不同(相关性要求).

第二步，在第一个主成分 t_1 和 u_1 被提取后，分别实施

① 各自变量对自变量系统第一主成分的回归(即用 t_1 表示 X).

② 各因变量对自变量系统第一主成分的回归(即用 t_1 表示 Y).

如果回归方程已经达到满意的精度，则算法终止；否则，将利用 X 被 t_1 解释后的残余信息以及 Y 被 t_1 解释后的残余信息进行第二轮的成分提取. 如此往复，直到能达到一个较满意的精度为止.

6.3.2 偏最小二乘回归方法的算法步骤

首先要进行预备分析，目的是判断自变量(因变量)是否存在多重相关性，判断因变量与自变量是否存在相关关系，进而决定是否需要采用 PLS 方法建模. 具体计算方法是：记矩阵 $Z = (X, Y)$，求 Z 的各列数据之间的简单相关系数；然后，按下列步骤建立偏最小二乘回归方程.

6.3.2.1　标准化原始数据

标准化后的数据矩阵记为 $\boldsymbol{E}_0=(e_{ij})_{n\times p}$ 和 $\boldsymbol{F}_0=(f_{ij})_{n\times q}$，其中

$$e_{ij}=\frac{x_{ij}-\bar{x}_j}{sx_j}\quad(i=1,2,\cdots,n;\quad j=1,2,\cdots,p),\tag{6.1}$$

$$f_{ij}=\frac{y_{ij}-\bar{y}_j}{sy_j}\quad(i=1,2,\cdots,n;\quad j=1,2,\cdots,q).\tag{6.2}$$

式(6.1)和(6.2)中，$\bar{x}_j$，$\bar{y}_j$ 分别为矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的第 j 列数据的平均值；sx_j，sy_j 分别为矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的第 j 列数据的标准差.

6.3.2.2　主成分提取

(1) 第一轮主成分提取

求矩阵 $\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0$ 的最大特征值所对应的单位特征向量 $\boldsymbol{w}_1$，得自变量的第 1 个主成分

$$\boldsymbol{t}_1=\boldsymbol{E}_0\boldsymbol{w}_1.$$

求矩阵 $\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0$ 的最大特征值所对应的单位特征向量 $\boldsymbol{c}_1$，得因变量的第 1 个主成分

$$\boldsymbol{u}_1=\boldsymbol{F}_0\boldsymbol{c}_1.$$

求残差矩阵

$$\boldsymbol{E}_1=\boldsymbol{E}_0-\boldsymbol{t}_1\boldsymbol{p}_1^{\mathrm{T}},\tag{6.3}$$

$$\boldsymbol{F}_1=\boldsymbol{F}_0-\boldsymbol{t}_1\boldsymbol{r}_1^{\mathrm{T}},\tag{6.4}$$

式(6.3)中 $\boldsymbol{p}_1=\dfrac{\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{t}_1}{\|\boldsymbol{t}_1\|^2}$，式(6.4)中 $\boldsymbol{r}_1=\dfrac{\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{t}_1}{\|\boldsymbol{t}_1\|^2}$.

在 PLS 方法中，称 $\boldsymbol{w}_1$ 为**模型效应权重**，$\boldsymbol{c}_1$ 为**因变量权重**，$\boldsymbol{p}_1$ 为**模型效应载荷量**.

(2) 新一轮主成分提取

令 $\boldsymbol{E}_0=\boldsymbol{E}_1$，$\boldsymbol{F}_0=\boldsymbol{F}_1$，回到(1)，对残差矩阵进行新一轮的主成分提取和回归分析.

设第 h 步的计算结果为

$$\boldsymbol{t}_h=\boldsymbol{E}_{h-1}\boldsymbol{w}_h,\tag{6.5}$$

$$\boldsymbol{u}_h=\boldsymbol{F}_{h-1}\boldsymbol{c}_h,\tag{6.6}$$

$$\boldsymbol{E}_h=\boldsymbol{E}_{h-1}-\boldsymbol{t}_h\boldsymbol{p}_h^{\mathrm{T}},\tag{6.7}$$

$$\boldsymbol{F}_h=\boldsymbol{F}_{h-1}-\boldsymbol{t}_h\boldsymbol{r}_h^{\mathrm{T}}.\tag{6.8}$$

式(6.5)至(6.8)中，$h=1,2,\cdots,m$，$m\leqslant\mathrm{rank}(\boldsymbol{E}_0)$，$\boldsymbol{p}_h=\dfrac{\boldsymbol{E}_{h-1}^{\mathrm{T}}\boldsymbol{t}_h}{\|\boldsymbol{t}_h\|^2}$，$\boldsymbol{r}_h=\dfrac{\boldsymbol{E}_{h-1}^{\mathrm{T}}\boldsymbol{t}_h}{\|\boldsymbol{t}_h\|^2}$.

(3) 主成分提取的终止准则

PLS 方法不需要选用所有的主成分建模，而是采用截尾的方式，即仅选择前 m 个主成分 $t_1,\cdots,t_m$，就可以得到一个预测性能较好的模型. 因此，在主成分提取的每一轮计算中，都要对是否得到了足够多的主成分进行判断.

判断准则常用的有交叉有效性准则和复测定系数准则.

定义 6.7（交叉有效性） 称

$$Q_h^2 = 1 - \frac{\text{PRESS}_h}{\text{SS}_{(h-1)}}$$

为主成分 t_h 关于因变量系统 $\boldsymbol{Y}$ 的**交叉有效性**.

上式中各参数的意义如下. PRESS_h 是从所有 n 个样本点中舍弃某个样本点 $x^{(i)}(i=1,2,\cdots,n)$之后，用剩余的 $n-1$ 个样本点拟合出含 h 个主成分的回归方程，再对 $x^{(i)}(i=1,2,\cdots,n)$点进行预测的预测误差平方和. 更详细一点，记 $\hat{y}_{hj(-i)}$ 为 y_j 在样本点 $x^{(i)}$ 上的预测值，$\text{PRESS}_{hj} = \sum_{i=1}^{n}[y_{ij} - \hat{y}_{hj(-i)}]^2$ 为 y_j 的预测误差平方和，则 $\text{PRESS}_h = \sum_{i=1}^{p}\text{PRESS}_{hi}$ 就是 $\boldsymbol{Y}$ 的**预测误差平方和**.

$\text{SS}_{(h-1)}$是用所有 n 个样本点拟合出的含 $h-1$ 个主成分的回归方程的拟合误差平方和. 更详细一点，记 $\hat{y}_{(h-1)ji}$ 为 y_j 在样本点 $x^{(i)}$ 上的拟合值，$\text{SS}_{(h-1)j} = \sum_{i=1}^{n}[y_{ij} - \hat{y}_{(h-1)ji}]^2$ 为 y_j 的拟合误差平方和，则 $\text{SS}_{(h-1)} = \sum_{i=1}^{p}\text{SS}_{(h-1)i}$ 就是 $\boldsymbol{Y}$ 的**拟合误差平方和**.

交叉有效性是对新增主成分能否对模型的预测功能有明显改进的判断指标.

若 $Q_h^2 \geqslant 1-0.95^2=0.0975$，则认为主成分 t_h 的边际贡献是显著的.

定义 6.8（复测定系数） 称

$$Q_h^2 = \frac{\sum_{k=1}^{h}(\|\boldsymbol{t}_k\|^2 \times \|\boldsymbol{p}_k\|^2)}{\|\boldsymbol{E}_0\|^2}$$

为自变量系统 $\boldsymbol{X}$ 被提取的**变异信息量**. 称

$$R_h^2 = \frac{\sum_{k=1}^{h}(\|\boldsymbol{t}_k\|^2 \times \|\boldsymbol{r}_k\|^2)}{\|\boldsymbol{F}_0\|^2}$$

为回归方程的**复测定系数**.

复测定系数表示所提取的主成分的可解释变异信息占总变异的百分比.

当 $h=m$，复测定系数 R_m^2 的值足够大时，可在第 m 步终止主成分的提取计算. 通

常 $R_m^2 \geq 0.85$ 即可.

6.3.2.3 建立回归方程

(1) 建立关于主成分的 MLR 回归方程

求出 $\boldsymbol{F}_0$ 在 $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_m$ 上的 MLR 回归方程

$$\boldsymbol{F}_0 = \boldsymbol{t}_1 \boldsymbol{r}_1^{\mathrm{T}} + \boldsymbol{t}_2 \boldsymbol{r}_2^{\mathrm{T}} + \cdots + \boldsymbol{t}_m \boldsymbol{r}_m^{\mathrm{T}} + \boldsymbol{F}_m. \tag{6.9}$$

(2) 变换为关于标准化变量的 PLS 回归方程

将 $\boldsymbol{t}_i = \boldsymbol{E}_{i-1} \boldsymbol{w}_i = \boldsymbol{E}_0 \boldsymbol{w}_i^*\ (i = 1, 2, \cdots, m)$ 代入方程(6.9)，得 $\boldsymbol{F}_0$ 关于 $\boldsymbol{E}_0$ 的 PLS 回归方程

$$\boldsymbol{F}_0 = \boldsymbol{E}_0 \boldsymbol{w}_1^* \boldsymbol{r}_1^{\mathrm{T}} + \boldsymbol{E}_0 \boldsymbol{w}_2^* \boldsymbol{r}_2^{\mathrm{T}} + \cdots + \boldsymbol{E}_0 \boldsymbol{w}_m^* \boldsymbol{r}_m^{\mathrm{T}} + \boldsymbol{F}_m. \tag{6.10}$$

其中 $\boldsymbol{w}_i^* = \prod_{k=1}^{i-1} (\boldsymbol{I} - \boldsymbol{w}_k \boldsymbol{p}'_k) \boldsymbol{w}_i (i = 1, 2, \cdots, m)$，$\boldsymbol{I}$ 为单位矩阵.

(3) 还原为关于原始变量的 PLS 回归方程

将方程(6.10)还原成关于原始变量的 PLS 回归方程

$$\hat{y}_k = \left(\bar{y}_k - \sum_{i=1}^{p} \alpha_{ki} \frac{sy_k}{sx_i} \bar{x}_i \right) + \sum_{i=1}^{p} \alpha_{ki} \frac{sy_k}{sx_i} x_i \quad (k = 1, 2, \cdots, q),$$

其中 $\boldsymbol{\alpha}_k$ 是矩阵 $\boldsymbol{\alpha}_{p \times q} = \sum_{j=1}^{m} \boldsymbol{w}_j^* \boldsymbol{r}'_j$ 的第 k 个列向量，α_{ki} 是 $\boldsymbol{\alpha}_k$ 的第 i 个分量.

6.3.3 偏最小二乘回归方法的辅助分析

PLS 方法除了前述建模技术，还包括 PLS 辅助分析技术，可以在获得一个更为合理的回归模型的同时，完成一些类似于主成分分析和典型相关分析的研究内容，提供更加丰富、深入的系统信息.

6.3.3.1 自变量和因变量之间的相关关系分析

在一元回归分析中，为了判定自变量和因变量之间的关系，经常采用散点图来作直观的分析，简单而有效. 这种方法在多元回归分析中遇到困难：多维数据构成了一个超平面，难以作直观观察；各自变量间相互关联，不能将变量简单地分割开来分析.

PLS 方法的 $\boldsymbol{t}_1/\boldsymbol{u}_1$ 平面图功能使这一点成为可能.

在 PLS 方法中，自变量集合 $\boldsymbol{X}$ 和因变量集合 $\boldsymbol{Y}$ 之间的相关关系可以通过 $\boldsymbol{t}_1$ 和 $\boldsymbol{u}_1$ 的相关关系得到反映. 因此，绘制以 $\boldsymbol{t}_1$ 为横坐标，$\boldsymbol{u}_1$ 为纵坐标的 $\boldsymbol{t}_1/\boldsymbol{u}_1$ 平面图，绘出第一主成分偶对($\boldsymbol{t}_1, \boldsymbol{u}_1$)的观测样本散点图. 如果所有样本点($t_1(i), u_1(i)$) ($i = 1, 2, \cdots, n$)在图中的排列近似于一条直线，则说明 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间存在着较强的相关关系，这时采用 PLS 方法建立 $\boldsymbol{Y}$ 对 $\boldsymbol{X}$ 的线性模型才会是合理的.

6.3.3.2　主成分对变量的解释能力的评价

在 PLS 计算过程中，要求所提取的自变量主成分 t_h 尽可能多地代表 X 的变异信息，尽可能与 Y 相关联，解释 Y 中的信息. 为了测量 t_h 对 X 和 Y 的解释能力，特给出如下定义.

定义 6.9（自变量的主成分对自变量系统的各种解释能力）

① 称主成分 t_h 与自变量 x_j 的简单相关系数的平方

$$\mathrm{Rd}(x_j;t_h)=r^2(x_j;t_h)$$

为 t_h 对某个自变量 x_j 的解释能力.

② 称

$$\mathrm{Rd}(X;t_h)=\frac{1}{p}\sum_{j=1}^{p}\mathrm{Rd}(x_j;t_h)$$

为 t_h 对自变量系统 X 的解释能力.

③ 称

$$\mathrm{Rd}(x_j;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(x_j;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对某个自变量 x_j 的累计解释能力.

④ 称

$$\mathrm{Rd}(X;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(X;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对自变量系统 X 的累计解释能力.

定义 6.10（自变量的主成分对因变量系统的各种解释能力）

① 称主成分 t_h 与因变量 y_j 的简单相关系数的平方

$$\mathrm{Rd}(y_j;t_h)=r^2(y_j;t_h)$$

为 t_h 对某个因变量 y_j 的解释能力.

② 称

$$\mathrm{Rd}(Y;t_h)=\frac{1}{q}\sum_{j=1}^{q}\mathrm{Rd}(y_j;t_h)$$

为 t_h 对因变量系统 Y 的解释能力.

③ 称

$$\mathrm{Rd}(y_k;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(y_k;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对某个因变量 y_k 的累计解释能力.

④ 称

$$\mathrm{Rd}(Y;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(Y;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对因变量系统 Y 的累计解释能力.

6.3.3.3 自变量对因变量系统的解释能力

PLS方法中，自变量对因变量的解释能力是以变量投影重要性指标(VIP)来测度的.

定义6.11（自变量对主成分的边际贡献） 称

$$\mathrm{VIP}_j=\sqrt{\frac{p}{\mathrm{Rd}(Y;t_1,\cdots,t_m)}\sum_{h=1}^{m}\mathrm{Rd}(Y;t_h)w_{hj}^2}$$

为自变量 x_j 对主成分 t_h 的边际贡献. 其中，w_{hj}是主轴 w_h 的第 j 个分量；$\mathrm{Rd}(Y;t_h)$，$\mathrm{Rd}(Y;t_1,t_2,\cdots,t_m)$分别是 t_h 对 Y 的解释能力和$t_1,t_2,\cdots,t_m$ 对 Y 的累计解释能力.

VIP_j 定义式的意义是基于这样一个事实：由于 x_j 对 Y 的解释是通过t_h 来传递的，如果 t_h 对 Y 的解释能力很强，而 x_j 在构造t_h 时又起到了相当重要的作用，则 x_j 对 Y 的解释能力就被视为很大. 也就是说，如果在$\mathrm{Rd}(Y;t_h)$值很大的 t_h 成分上，w_{hj}取很大的值，则 x_j 对解释 Y 就有很重要的作用.

另外，容易证明 $\sum_{j=1}^{p}\mathrm{VIP}_j^2=p$，所以，对于 p 个自变量$x_j(j=1,2,\cdots,p)$，如果它们在解释 Y 时的作用都相同，则所有 VIP_j 均等于1；否则，对于 $\mathrm{VIP}_j(>1)$很大的 x_j，它在解释因变量 Y 时就有更加重要的作用.

希望深入了解 PLS 建模理论与方法的读者可参阅文献[12]和[13].

【例6.5】 为研究辽宁省教育投入与产业发展之间的相关关系，收集了如表6.6所示的数据资料.

表6.6　辽宁省1984—2005年教育投入与经济产出数据资料

年份	L_1	L_2	L_3	L_4	K	Y_1	Y_2	Y_3
1984	122	15612	564419	512965	73961	80.4	268.2	89.6
1985	584	17495	522327	689598	102450	74.9	328.1	115.6
1986	670	20583	517410	704016	123383	92.9	357.8	154.6
1987	1193	29394	549709	680861	124532	109.5	417.0	192.6
1988	1929	31552	615839	637753	155617	141.9	492.5	246.6
1989	1763	32708	598834	593257	194395	141.9	545.1	316.9
1990	1677	33768	580075	591654	201077	168.6	540.8	353.3
1991	1500	33530	571569	660343	229033	180.8	590.1	429.2
1992	1245	35208	573509	685996	254712	194.6	741.9	536.5
1993	1307	33615	572612	630759	305120	260.8	1039.3	710.8
1994	1273	35923	606148	636786	398399	319.0	1259.1	883.8
1995	1425	44072	635387	672482	439517	392.2	1390.0	1011.2

续表 6.6

年份	L_1	L_2	L_3	L_4	K	Y_1	Y_2	Y_3
1996	1962	51068	611379	576164	496190	474.1	1537.7	1145.9
1997	2316	49591	666386	500252	546883	474.1	1743.9	1364.2
1998	2126	47557	724391	555892	562770	531.5	1855.2	1459.1
1999	2426	49964	658165	644042	642559	520.8	2001.5	1649.4
2000	2910	49834	587000	722325	760719	503.4	2344.4	1821.2
2001	2971	60271	623975	679852	855043	544.4	2440.6	2048.1
2002	3674	72791	709233	622536	991450	590.2	2609.9	2258.2
2003	5027	98908	788473	595278	1108785	615.8	2898.9	2487.9
2004	6726	115889	792228	511757	1387080	798.4	3061.6	2812.0
2005	9342	144984	815905	499069	1629956	882.4	3953.3	3173.3

表 6.6 中数据摘自《辽宁统计年鉴 2006》，各变量的意义及数量单位如下：

① 教育投入水平的指标：

L_1——研究生教育程度(硕士及博士)劳动力数(单位：人)；

L_2——高等教育程度(大学本科及专科)劳动力数(单位：人)；

L_3——中等教育程度(高中及中专)劳动力数(单位：人)；

L_4——初等以下教育程度(小学及文盲)劳动力数(单位：人)；

K——教育的财政投入(单位：万元).

② 经济产出的指标：

Y_1——第一产业(包括林业、牧业、渔业等)产出值(单位：亿元)；

Y_2——第二产业(包括工业和建筑业)产出值(单位：亿元)；

Y_3——第三产业(包括流通类的交通运输业、邮电通讯业、商业饮食业、物资供销和仓储业及金融、保险业、地质普查业、房地产、公用事业、居民服务业、旅游业、咨询信息服务业和各类技术服务业，等等)产出值(单位：亿元).

建模分析如下.

(1) 多重相关性诊断

① 计算自变量与因变量之间的相关系数.

```
load jytrjjcc % 预先编写数据文件 jytrjjcc.mat，并保存到当前工作路径下
cr = corrcoef(jytrjjcc); % 计算变量之间的相关系数
```

计算结果整理见表 6.7.

表 6.7　　因变量与自变量之间的相关系数

r	L_1	L_2	L_3	L_4	K	Y_1	Y_2	Y_3
L_1	1.0000	0.9847	0.8737	−0.4847	0.9447	0.8643	0.8906	0.8895
L_2		1.0000	0.9117	−0.4944	0.9695	0.9088	0.9250	0.9278
L_3			1.0000	−0.6196	0.8944	0.8940	0.8776	0.8870
L_4				1.0000	−0.4177	−0.4436	−0.3751	−0.3803
K					1.0000	0.9635	0.9833	0.9871
Y_1						1.0000	0.9827	0.9818
Y_2							1.0000	0.9961
Y_3								1.0000

由表 6.7 可以看出：自变量之间的相关系数最高达 0.9847，表明自变量之间存在严重的自相关性. 注意，初等以下教育程度劳动力数与其他自变量之间呈负相关关系.

因变量与自变量之间的相关系数最高达 0.9871，表明自变量系统与因变量系统之间存在较高的相关性. 注意：研究生和高等、中等教育程度劳动力数以及财政投入与三大产业产出之间存在着明显的正相关关系，而初等以下教育程度劳动力数与三大产业产出之间存在着较弱的负相关关系.

② 建立普通最小二乘回归方程.

原始数据标准化，得到自变量的标准化数据矩阵 E_0 和因变量的标准化数据矩阵 F_0，再建立两者之间的多重(多因变量)多元线性回归方程(MLR).

```
E0 = stand(jytrjjcc(: , 1: 5)); %标准化自变量数据
F0 = stand(jytrjjcc(: , 6: 8)); %标准化因变量数据
MMLR = inv((E0′ * E0)) * (E0′ * F0); %估计多重多元线性回归方程系数
```

根据上述计算结果，可得下列多重多元线性回归方程：

$$F_{01} = -0.4171E_{01} - 0.1685E_{02} + 0.1873E_{03} - 0.0578E_{04} + 1.32924E_{05},$$

$$F_{02} = -0.2410E_{01} - 0.1647E_{02} - 0.0209E_{03} - 0.0071E_{04} + 1.3867E_{05},$$

$$F_{03} = -0.2237E_{01} - 0.3039E_{02} + 0.1530E_{03} + 0.0270E_{04} + 1.3674E_{05}.$$

从这一组回归方程可以看出，三大产业产出值与研究生教育、高等教育竟然负相关，这与客观事实相违背，也与相关系数矩阵中得到的结论相悖.

所以，在自变量之间以及自变量与因变量之间存在复杂的相关关系时，普通最小二乘回归方法建立的模型不能准确地反映实际情况. 这种情况下可采用偏最小二乘回归分析方法建模.

(2) 建立偏最小二乘回归模型

① 提取所有可能的主成分.

```
clear
load jytrjjcc
X = jytrjjcc(:,1:5);
Y = jytrjjcc(:,6:8);
E0 = stand(X);
F0 = stand(Y);
A = rank(E0);
[W,C,T,U,P,R] = plspcr(E0,F0); % 提取所有可能的主成分
```

② 主成分解释能力分析.

首先，计算主成分累积复测定系数.

```
RA = plsra(T,R,F0,A)
```

上述指令的运行结果是：

```
RA =
    0.8727    0.9209    0.9739    0.9870    0.9879
```

计算结果表明：抽取一个主成分时，回归方程的复测定系数已达到 87.27%；抽取两个主成分时，回归方程的复测定系数已达到 92.09%；等等. 通常，系统信息的可解释变异达到总变异的 85% 即可认为回归方程的精度已达到满意效果. 因此，根据模型从简的原则，我们只需选取一个主成分建模. 第一主成分的表达式为

$$t_1 = E_0 w_1 = -0.4694E_{01} - 0.4902E_{02} - 0.4719E_{03} + 0.2128E_{04} - 0.5208E_{05}.$$

接下来计算主成分的信息解释能力.

```
[Rdx,RdX,RdXt,Rdy,RdY,RdYt] = plsrd(E0,F0,T,A)
```

上述指令的运行结果是：

```
Rdx =
    0.9421    0.0092    0.0444    0.0017    0.0025
    0.9744    0.0110    0.0083    0.0006    0.0057
    0.9108    0.0054    0.0306    0.0530    0.0003
    0.3490    0.6425    0.0046    0.0040    0.0000
    0.9335    0.0433    0.0080    0.0150    0.0001
RdX =
    0.8220    0.1423    0.0192    0.0149    0.0017
RdXt =
    1.0000
Rdy =
```

```
    0.8573    0.0252    0.0670    0.0104    0.0001
    0.8650    0.0597    0.0415    0.0153    0.0018
    0.8728    0.0584    0.0493    0.0130    0.0009
RdY =
    0.8650    0.0478    0.0526    0.0129    0.0009
RdYt =
    0.9793
```

对上述计算结果进行简化和整理，见表6.8.

表6.8 主成分 t_1 和 t_2 对变量的解释能力

Rd	L_1	L_2	L_3	L_4	K	Y_1	Y_2	Y_3	X	Y
t_1	0.9421	0.9744	0.9108	0.3490	0.9335	0.8573	0.8650	0.8728	0.8220	0.8650
t_2	0.0092	0.0110	0.0054	0.6425	0.0433	0.0252	0.0597	0.0584	0.1423	0.0478

从表6.8中可以看出，主成分 t_1 对变量 L_1,L_2，L_3 和 K 的解释能力均相当强，而对 L_4 的解释能力较弱，因此可以认为 t_1 是由变量 L_1,L_2,L_3 和 K 综合而成的，并且解释了原自变量系统82.20%的变异信息，对原自变量系统有非常好的代表性. 同时，解释了因变量系统86.50%的信息，对因变量系统的贡献很大. 而第二个主成分 t_2 主要代表的是变量 L_4，对原自(因)变量系统信息变异的解释能力较低.

经计算，当增加第二个主成分 t_2 时，模型的精度没有明显的改善. 因此，从主成分的信息解释能力的角度以及模型从简的原则，只选一个主成分建模是适宜的.

③ 考查第一主成分间的相关性.

绘制 t_1/u_1 图直观地考查第一主成分间的相关性.

```
cr = plsutcor(U,T)
```

上述指令的运行结果是:

```
cr =
   1.0000    0.9342
   0.9342    1.0000
```

从图6.3中可以看出，自变量系统与因变量系统第一主成分间的相关性很强，适合建立线性回归模型.

④ 求PLS回归方程的系数.

先求标准化因变量 F_{01}，F_{02}，F_{03} 关于主成分 t_1 的经验回归系数.

```
TCOEFF = R(:, 1)%这组系数存于 plspcr 函数的最后一个输出变量 R 中
```

上述指令的运行结果是:

```
TCOEFF =
```

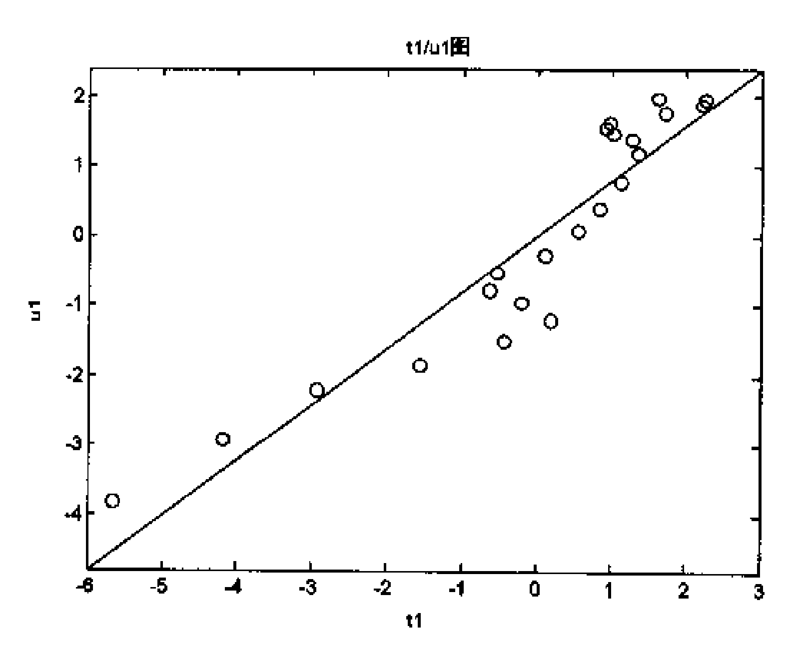


图 6.3 t_1/u_1 散点图

```
  -0.4586
  -0.4607
  -0.4627
```

再求标准化变量 F_{01}，F_{02}，F_{03}关于 E_{01}，E_{02}，E_{03}，E_{04}，E_{05}的经验回归系数.

```
SCOEFF = pls(1,5,W,P,R)
```

上述指令的运行结果是：

```
SCOEFF =
     0.2153      0.2163      0.2172
     0.2248      0.2258      0.2269
     0.2164      0.2174      0.2184
    -0.0976     -0.0980     -0.0985
     0.2389      0.2399      0.2410
```

最后求原始变量 Y_1，Y_2，Y_3 关于 L_1，L_2，L_3，L_4，K 的经验回归系数.

```
[COEFF,INTERCEP] = plsiscoeff(X,Y,SCOEFF)
```

上述指令的运行结果是：

```
COEFF =
     0.0242      0.1072      0.0966
     0.0017      0.0074      0.0067
     0.0006      0.0027      0.0024
    -0.0003     -0.0015     -0.0014
```

```
       0.0001        0.0006        0.0005
INTERCEP =
     -17.9677     -233.0059     -388.8328
```

根据上述计算结果，写出各个阶段所建回归方程如下.

F_0 关于成分 t_1 的(MLR)回归方程为

$$F_{01} \approx r_{11}t_1 = -0.4586t_1,$$

$$F_{02} \approx r_{12}t_2 = -0.4607t_1,$$

$$F_{03} \approx r_{13}t_3 = -0.4627t_1,$$

将 $t_1 = E_0 w_1 = -0.4694E_{01} - 0.4902E_{02} - 0.4719E_{03} + 0.2128E_{04} - 0.5208E_{05}$ 代入上面的三个方程，得 F_0 关于 E_0 的(PLS)回归方程为

$$F_{01} \approx 0.2153E_{01} + 0.2248E_{02} + 0.2164E_{03} - 0.0976E_{04} + 0.2389E_{05},$$

$$F_{02} \approx 0.2163E_{01} + 0.2258E_{02} + 0.2174E_{03} - 0.0980E_{04} + 0.2399E_{05},$$

$$F_{03} \approx 0.2172E_{01} + 0.2269E_{02} + 0.2184E_{03} - 0.0985E_{04} + 0.2410E_{05}.$$

由逆标准化变换，将上述三个方程还原为原始因变量关于自变量的(PLS)回归方程为

$$\hat{Y}_1 = -17.9677 + 0.0242L_1 + 0.0017L_2 + 0.0006L_3 - 0.0003L_4 + 0.0001K,$$

$$\hat{Y}_2 = -233.0059 + 0.1072L_1 + 0.0074L_2 + 0.0027L_3 - 0.0015L_4 + 0.0006K,$$

$$\hat{Y}_3 = -388.8328 + 0.0966L_1 + 0.0067L_2 + 0.0024L_3 - 0.0014L_4 + 0.0005K.$$

可见，所建的回归方程没有出现反符号现象，受中等以上教育的劳动力人数、财政投入与经济的产出都是呈正相关的，只有初等教育劳动力人数(包括文盲)呈负相关，这与相关系数符号完全一致.

(3) 变量投影重要性分析与模型的改进

下面从变量投影重要性的角度分析回归方程中自变量对因变量的解释能力.

```
VIP = plsvip(W, RdY, RdYt, 1)
```

上述指令的运行结果是:

```
VIP =
    0.9866      1.0303      0.9918      0.4472      1.0946
```

变量投影重要性指标是用来测度第 j 个自变量对因变量的解释能力的. 因此，从预测的角度，如果某个自变量在解释因变量时起的作用很小，则可以考虑章去这个变量后重新建模. 由图 6.4 可以看出，VIP_4 明显较小，故删除变量 L_4 重新用偏最小二乘回归方法建模，得到的回归方程为

$$\hat{Y}_1 = -270.7 + 0.0260L_1 + 0.0018L_2 + 0.0006L_3 + 0.0001K,$$

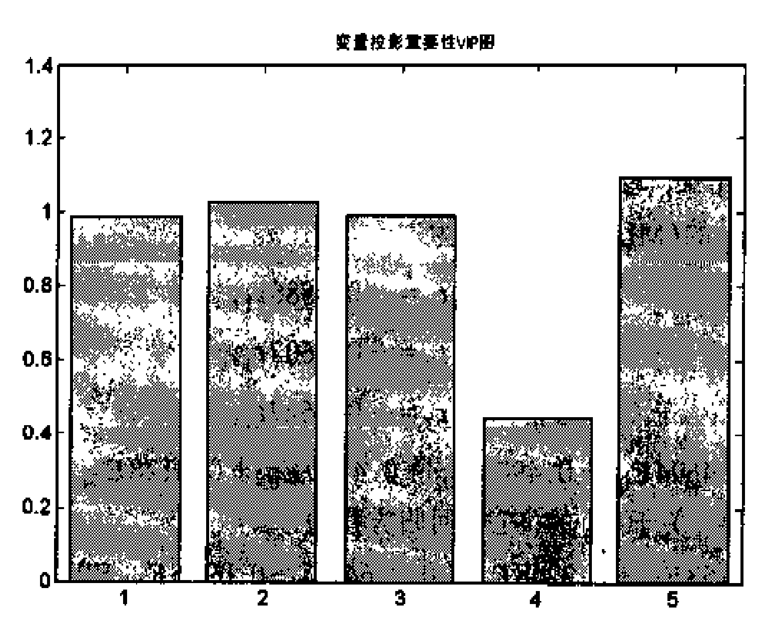


图 6.4　变量投影重要性 VIP 图

$$\hat{Y}_2 = -1374.79 + 0.1158L_1 + 0.0080L_2 + 0.0029L_3 + 0.0006K,$$

$$\hat{Y}_3 = -1416.1 + 0.1043L_1 + 0.0072L_2 + 0.0026L_3 + 0.0005K.$$

与未删除变量 L_4 前的回归方程对比，发现方程的回归系数变化很小．深入的精度分析结果见表 6.9.

表 6.9　改进前后模型应用效果比对分析

	Rdx	Rdy	SS	PRESS
包含 L_4 的模型	0.8220	0.8650	8.5034	8.3149
删除 L_4 的模型	0.9476	0.8893	6.9714	6.4205

表 6.9 中，SS 值表示的是回归方程对所有样本点的拟合误差平方和，PRESS 值表示的是预测误差平方和，计算公式详见文献[13]．由表 6.9 可知，删除变量 L_4 后的模型，无论是建模的主成分 t_1 对自(因)变量的解释能力 Rdx(Rdy)，还是拟合与预测效果上都有很明显的提高．因此，基于 VIP 对自变量筛选后的偏最小二乘回归模型效果更佳.

需要强调的是，删除 L_4 的模型对分析教育投入与经济产出两者之间关系来说意义并不是很大，但若考虑对辽宁省经济产出进行短期预测，采用该模型的预测精度会更高.

对上述统计分析信息的深入解读依赖更多的教育经济学领域的专业知识和经验，已超出本书的范畴.

本例数据处理所用 MATLAB 偏最小二乘回归建模函数均为自定义 M-函数，函数的源代码见本书附录 C.

习题 6

1. 表 6.10 数据是退火温度 x(单位：℃)对黄铜延性 y 效应的试验结果，y 是以延伸率计算的，且设为正态变量，求 y 对 x 的样本回归方程.

表 6.10

x/℃	300	400	500	600	700	800
y/%	40	50	55	60	67	70

2. 某健身俱乐部对部分会员进行了一项调查，现将会员的入会时间(Y)和到俱乐部的次数(X)统计如下，见表 6.11.

表 6.11

入会时间/月	12	2	6	9	7	2	8	4	10	5
健身次数	4	10	8	5	5	8	3	8	2	5

试完成下列问题：

(1) 画出散点图；

(2) 建立适当的回归方程；

(3) 对方程进行检验.

3. 某公司在 15 个地区的某种商品的销售量 y(单位：罗，1 罗 = 12 打)和各地区人口数 x_1(单位：千人)，以及平均每户总收入数 x_2(单位：元)的统计资料见表 6.12.

表 6.12

地区	x_{1i}	x_{2i}	y_i	地区	x_{1i}	x_{2i}	y_i
1	274	2450	162	9	195	2137	116
2	180	3254	120	10	53	2560	55
3	375	3802	223	11	430	4020	252
4	205	2838	131	12	372	4427	232
5	86	2347	67	13	236	2660	144
6	265	3782	169	14	157	2088	103
7	98	3008	81	15	370	2605	212
8	330	2450	192				

求 y 对 x_1，x_2 的回归方程，并根据人口数、每户总收入数预测某地区的销售量.

4. 10 个同类企业的生产性固定资产价值和工业总产值资料如表 6.13 所示.

表 6.13

企业编号	生产性固定资产价值/万元	工业总产值/万元
1	318	524
2	910	1019
3	200	638
4	409	815
5	415	913
6	502	928
7	314	605
8	1210	1516
9	1022	1219
10	1225	1624
合计	6525	9801

试完成下列问题：

(1) 说明两变量之间的相关方向；

(2) 建立回归直线方程；

(3) 计算估计标准差；

(4) 估计生产性固定资产价值(自变量)为 1100 万元时总产值(因变量)的可能值.

5. 某公司采集了市场上办公用房的空闲率和租金率的数据. 表 6.14 是选取的 18 个城市中心商业区的综合空闲率(单位：%)和平均租金率(单位：元/m^2)的数据.

表 6.14

地区编号	综合空闲率	平均租金率	地区编号	综合空闲率	平均租金率
1	21.9	18.54	10	6.6	31.42
2	6.0	33.70	11	15.9	18.74
3	22.8	19.67	12	9.2	26.76
4	18.1	21.01	13	19.7	27.72
5	12.7	35.09	14	20.0	18.20
6	14.5	19.41	15	8.3	25.00
7	20.0	25.28	16	17.1	29.78
8	19.2	17.02	17	10.8	37.03
9	16.0	24.04	18	11.1	28.64

试完成下列问题：

(1) 用横轴表示空闲率，对这些数据画出散点图；

(2) 这两个变量之间能显示出什么关系吗?

(3) 在办公用房的综合空闲率已知时，求出能用来预测平均租金率的回归方程；

(4) 在0.05显著性水平下检验关系的显著性；

(5) 估计的回归方程对数据的拟合好吗？请作出解释；

(6) 在一个综合空闲率是25%的中心商业区，预测该市场的期望租金率；

(7) 若某市的中心商业区综合空闲率是11.3%，预测该市中心商业区的期望租金率.

6. 某公司的管理者认为每周的收入是广告费用的函数，并想对每周的总收入(单位:千元)作出估计. 由8周的历史数据组成的样本如表6.15所示.

表6.15

每周的总收入	电视广告费用/千元	报纸广告费用/千元
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.5
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

试完成下列问题：

(1) 将电视广告费用作为自变量，建立回归方程.

(2) 将电视广告费用与报纸广告费用作为自变量，建立回归方程.

(3) 在上面建立的估计的回归方程中，电视广告费用的系数相同吗？对每一种情形的系数作出解释.

(4) 若电视广告费用为3500元，报纸广告费用为1800元，一周总收入的估计值是多少？

(5) 对于模型 $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$，在0.05显著性水平下，检验假设 $H_0:\beta_1=\beta_2=0$. 其中：x_1 为电视广告费用(单位:千元)，x_2 为报纸广告费用(单位:千元).

(6) 在0.05显著性水平下，检验 β_1 的显著性，x_1 应该从模型中删去吗？

(7) 在0.05显著性水平下，检验 β_2 的显著性，x_2 应该从模型中删去吗？

7. "飞鸽"公司是一家生产自行车和与自行车相关零部件的企业，管理人员认为自行车的销售量(单位：千辆)依赖于本公司自行车的价格及其竞争对手的价格(单位：元)，并希望建立自行车的销售量与该公司自行车价格和竞争厂商自行车价格的回归方程. 表6.16列出了10个城市的价格资料.

表 6.16

竞争厂商的价格 x_1	该公司的价格 x_2	销售数量 y
256	240	102
280	260	100
380	352	120
260	300	77
310	320	46
350	300	93
250	300	26
290	300	69
360	400	65
300	350	85

试完成下列问题：

(1) 建立回归方程，要求它能在竞争厂商自行车的价格与该公司自行车的价格已知时预测自行车的销售量；

(2) 对回归方程中的 b_1 和 b_2 作出解释；

(3) 如果在一个城市"飞鸽"自行车的销售价格为 270 元，竞争厂商的自行车价格为 280 元，预测在该城市自行车的销售量.

8. 某电器经销公司在 15 个城市设有经销处，公司发现彩电销售量与该城市居民数多少有关系，并希望通过居民数的多少来预测其彩电销售量. 表 6.17 是有关彩电销售量与城市居民户数的统计数据.

表 6.17

城市编号	销售量/台	户数/万户	城市编号	销售量/台	户数/万户
1	5425	189	9	5375	182
2	6319	193	10	4500	175
3	6827	197	11	3310	161
4	7743	202	12	8239	214
5	8365	206	13	4596	166
6	8916	209	14	3652	163
7	5970	185	15	4203	167
8	4719	179			

试完成下列问题：

(1) 计算彩电销售量与城市居民之间的线性相关系数；

(2) 拟合彩电销售量对城市居民户数的回归直线;

(3) 计算决定系数;

(4) 对回归方程的线性关系和回归系数进行显著性检验($\alpha=0.05$),并对结果作简要分析.

9. 某种商品销售量(Y),消费者的平均收入(X_1)以及商品价格(X_2)的统计数据如表6.18所示.

表6.18

Y	100	75	80	70	50	60	110	100	90	65
X_1	1000	600	1200	500	300	300	1300	1100	1300	400
X_2	5	7	6	6	8	9	3	4	5	7

试完成下列问题:

(1) 建立 Y 对 X_1, X_2 的线性回归方程;

(2) 对方程进行显著性检验;

(3) 对偏回归系数进行检验;

(4) 当 $X_1=1200$, $X_2=8$ 时,在95%的置信度下,求销售量 Y 的置信区间.

10. 某矿区采取18个煤样,测得密度及灰分(单位:%)见表6.19.

表6.19

样品	密度	灰分	样品	密度	灰分	样品	密度	灰分
1	1.5	25	7	1.3	5	13	1.6	25
2	1.2	4	8	1.5	24	14	1.4	6
3	1.7	30	9	1.7	33	15	1.6	26
4	1.4	20	10	1.3	4	16	1.5	24
5	1.8	36	11	1.5	17	17	1.4	20
6	1.3	7	12	1.5	24	18	1.4	9

试求密度及灰分之间的线性回归方程,并在显著性水平0.01下检验其线性相关程度.

11. 为研究学习时间长短对某门功课学习成绩的影响,现随机抽取20个学生,得到如表6.20所示的资料.

表 6.20

编号	学习时间/学时	成绩分数	编号	学习时间/学时	成绩分数
1	40	40	11	90	80
2	40	60	12	90	85
3	50	60	13	95	85
4	60	65	14	95	90
5	65	70	15	95	92
6	70	75	16	100	92
7	70	78	17	100	90
8	80	78	18	100	85
9	85	80	19	110	95
10	85	80	20	110	90

试完成下列问题：

(1) 判断学习时间长短与学习成绩之间有无线性相关关系；

(2) 在显著性水平为 5% 时，检验学习时间长短与学习成绩之间的线性相关程度是否显著；

(3) 若有显著性的线性相关关系，求出两者之间的线性回归方程，指出学习时间为 100 学时时成绩的平均数；

(4) 在显著性水平为 0.05 时，对回归参数进行统计检验；

(5) 计算估计标准误差.

12. 合成纤维的强度 η(单位:kg/mm^2)与其拉伸倍数 x 有关，测得试验数据见表 6.21.

表 6.21

x_i	2.0	2.5	2.7	3.5	4.0	4.5	5.2	6.3	7.1	8.0	9.0	10.0
η_i	1.3	2.5	2.5	2.7	3.5	4.2	5.0	6.4	6.3	7.0	8.0	8.1

试完成下列问题：

(1) 求 η 对 x 的回归直线；

(2) 在显著性水平为 0.05 时检验回归直线的显著性；

(3) 求 $x_0=6$ 时，η_0 的预测值及预测区间(置信度为 0.95).

附录 A　MATLAB 的基本函数

一、通用命令

(1) 通用信息

函数	功能
help	在线帮助

(2) 工作空间管理

函数	功能
clear	清除内存变量和函数
who	列出内存中的变量名
whos	列出内存中变量的详细信息
pack	收集 MATLAB 内存碎块扩大内存
save	把内存变量保存为文件
load	从 MAT 文件读取变量
quit	退出 MATLAB 环境
exit	退出 MATLAB 环境

(3) 函数管理

函数	功能
what	列出当前目录上的文件
which	确定函数、文件的位置
type	显示 M-文件
lookfor	按关键字搜索 M-文件
inmem	列出内存中的函数名

(4) 命令窗口控制与操作系统命令

函数	功能
cd	指定当前目录
clc	清除指令窗
diary	MATLAB 指令窗文本内容记录
dir	目录列表

dos	执行 DOS 指令并返回结果
echo	M-文件被执行指令的显示
format	设置输出格式
more	指令窗中内容的分页显示

二、基本数学函数

(1) 三角函数

函数	功能
sin	正弦
sinh	双曲正弦
cos	余弦
cosh	双曲余弦
tan	正切
tanh	双曲正切
asin	反正弦
asinh	反双曲正弦
acos	反余弦
acosh	反双曲余弦
atan	反正切
atanh	反双曲正切
atan2	四象限反正切

(2) 指数函数

函数	功能
exp	指数函数
pow2	2 的幂
log	自然对数
log2	底为 2 的对数
sqrt	平方根

(3) 复数函数

函数	功能
abs	绝对值、模、字符的 ASCⅡ码值
conj	复数共轭
real	复数的实部

imag	复数的虚部

(4) 数据分析与其他数学函数

函数	功能
min	找向量中最小元素
max	找向量中最大元素
rem	求余数
sign	符号函数
cumsum	元素累计和
sum	元素和
diff	数值差分、符号微分
int	符号积分
fft	离散 Fourier 变换
fftn	高维离散 Fourier 变换
expand	符号计算中的展开操作

(5) 数值处理函数

函数	功能
round	向最近整数取整
fix	向零取整
ceil	向正无穷取整
floor	向负无穷取整

三、矩阵与数值线性代数

(1) 特殊变量与常数

函数	功能
eps	浮点相对精度
i,j	虚数单位
pi	圆周率
inf	无穷大
NaN	不定式(非数)变量

(2) 基本矩阵生成函数

函数	功能
eye	单位阵

ones	全 1 数组
zeros	全 0 数组
rand	产生均匀分布随机数
randn	产生正态分布随机数

(3) 矩阵操作

函数	功能
tndims	求数组维数
reshape	改变数组维数、大小
length	数组长度
size	矩阵的大小
find	寻找非零元素下标
end	数组每维最后元素下标
ril	下三角阵
triu	上三角阵
full	把稀疏矩阵转换为非稀疏阵

(4) 高维数组与其他数据类型的创建与操作

函数	功能
cat	串接成高维数组
cell	创建元胞数组
struct	创建构架数组
fieldnames	构架域名
cell2struct	把元胞数组转换为构架数组
struct2cell	把构架数组转换为元胞数组
char	把数值、符号、内联类对象转换为字符对象

(5) 特殊矩阵

函数	功能
diag	矩阵对角元素提取、创建对角阵
magic	魔方阵
sparse	创建稀疏矩阵

(6) 矩阵函数与线性方程组

函数	功能
det	行列式

inv	求矩阵逆
norm	矩阵或向量范数
lu	LU 分解
svd	奇异值分解
chol	Cholesky 分解
eig	求特征值和特征向量
expm	常用矩阵指数函数
permute	广义转置
rcond	矩阵倒条件数估计

四、程序设计语言与调试

(1) 程序控制流程

函数	功能
if	条件分支结构
switch	多分支结构
for	构成 for 循环
while	控制流中的 While 循环结构
end	控制流 for 等结构体的结尾
try	控制流中的 Try-catch 结构
break	while 或 for 环中断指令
eturn	返回调用函数
errortrap	错误发生后程序是否继续执行的控制

(2) 变量、赋值与执行

函数	功能
assignin	向变量赋值
global	定义全局变量
double	把其他类型对象转换为双精度数值

(3) 程序参数处理

函数	功能
inputname	输入宗量名
nargin	函数输入宗量数
nargout	函数输出宗量数

(4) 信息显示

函数	功能
fprintf	设置显示格式
disp	显示数组
lasterr	显示最新出错信息
lastwarn	显示最新警告信息
warning	显示警告信息
error	显示出错信息并中断执行
display	显示对象内容的重载函数

(5) 交互输入

函数	功能
input	提示用户输入
keyboard	键盘获得控制权
pause	暂停

(6) 其他

函数	功能
class	获知对象类别或创建对象
methods	获知对指定类定义的所有方法函数
superiorto	设定优先级
flops	打开外部文件
fread	从文件读二进制数据
clock	时钟
drawnow	更新事件队列，强迫 MATLAB 刷新屏幕

五、绘图与图形界面设计

(1) 基本绘图函数

函数	功能
image	显示图像
plot	二维线图
plot3	三维线图
fill	二维多边形填色图
fill3	三维多边形填色图
surf	三维着色表面图
rectangle	画长方框

surfc	带等位线的表面图
text	文字注释
semilogx	X 轴对数刻度坐标图
semilogy	Y 轴对数刻度坐标图
loglog	双对数刻度图形

(2) 用户图形界面设计

函数	功能
get	获知对象属性
set	设置图形对象属性
findobj	寻找具有指定属性的对象图柄
cdedit	启动用户菜单、控件回调函数设计工具
axes	创建轴对象的低层指令
uicontextmenu	创建现场菜单
uicontrol	创建用户控件
uimenu	创建用户菜单
patch	创建块对象
figure	创建图形窗
light	创建光对象
line	创建线对象

(3) 动画设计

函数	功能
movie	放映影片动画
getframe	获取影片的帧画面

六、字符串处理

函数	功能
sprintf	把格式数据写成串
sscanf	按指定格式读串
strcmp	串比较
strncmp	串中前若干字符比较
strrep	串替换
findstr	寻找短串的起始字符下标
lower	转换为小写字母

upper	转换为大写字母
feval	执行由串指定的函数
eval	串演算指令
evalin	跨空间串演算指令

七、逻辑判断与检测

函数	功能
exist	检查变量或函数是否已定义
isa	检测是否给定类的对象
all	所有元素非零为真
any	所有元素非全零为真
isreal	若是实数则为真
isequal	若两数组相同则为真
isempty	若是空阵则为真
isfinite	若全部元素都有限则为真
islogical	若是逻辑数组则为真
isinf	若是无穷数据则为真
isnan	若是非数则为真
issparse	若是稀疏矩阵则为真
ischar	若是字符串则为真
isglobal	若是全局变量则为真
isletter	若是英文字母则为真
isspace	若是空格则为真
ishandle	若是图形句柄则为真

八、其他

函数	功能
sim	运行 SIMULINK 模型
simset	对 SIMULINK 模型的仿真参数进行设置
simulink	启动 SIMULINK 模块库浏览器

附录 B　MATLAB 常用统计分析函数

一、关于概率分布的 MATLAB 描述

(1) 20 种常见分布的 MATLAB 名称

分布类型	分布的 MATLAB 名称
贝塔分布	beta 或 Beta
伽玛分布	gam 或 Gamma
指数分布	exp 或 Exponential
正态分布	norm 或 Normal
对数正态分布	logn 或 Lognormal
均匀分布	unif 或 Uniform
瑞利分布	rayl 或 Rayleigh
威布尔分布	weib 或 Weibull
二项分布	bino 或 Binomial
泊松分布	poiss 或 Poisson
几何分布	geo 或 Geometric
超几何分布	hyge 或 Hypergeometric
离散均匀分布	unid 或 Discrete Uniform
负二项式分布	nbin 或 Negative Binomial
卡方分布	chi2 或 Chisquare
T 分布	t 或 T
F 分布	f 或 F
非中心卡方分布	ncx2 或 Noncentral Chisquare
非中心 F 分布	ncf 或 Noncentral F
非中心 t 分布	nct 或 Noncentral t

(2) 概率密度函数(pdf)

函数及其调用格式	功能与参数说明
y=betapdf(x,a,b)	求参数为 a,b 的 β 分布在 x 处的概率密度值 y
y=gampdf(x,a,b)	求参数为 a,b 的 γ 分布在 x 处的概率密度值 y

y=exppdf(x,lambda)	求参数为 lambda 的指数分布在 x 处的概率密度值 y
y=normpdf(x,mu,sigma)	求参数为 mu,sigma 的正态分布在 x 处的概率密度值 y
y=lognpdf(x,mu,sigma)	求参数为 mu,sigma 的对数正态分布在 x 处的概率密度值 y
y=unifpdf (x,a,b)	求区间[a,b]上的均匀分布在 x 处的概率密度值 y
y=raylpdf(x,b)	求参数为 b 的瑞利分布在 x 处的概率密度值 y
y=weibpdf(x,a,b)	求参数为 a,b 的威布尔分布在 x 处的概率密度值 y
y=binopdf(x,n,P)	求参数为 n,P 的二项分布在 x 处的概率密度值 y
y=poisspdf(x,lambda)	求参数为 lambda 的泊松分布在 x 处的概率密度值 y
y=geopdf(x,P)	求参数为 P 的几何分布在 x 处的概率密度值 y
y=hygepdf(x,m,k,n)	求参数为 m,k,n 的超几何分布在 x 处的概率密度值 y
y=unidpdf(x,n)	求参数为 n 的离散均匀分布在 x 处的概率密度值 y
y=nbinpdf(x,R,P)	求参数为 R,P 的负二项式分布在 x 处的概率密度值 y
y=chi2pdf(x,n)	求自由度为 n 的卡方分布在 x 处的概率密度值 y
y=tpdf(x,n)	求自由度为 n 的 t 分布在 x 处的概率密度值 y
y=fpdf(x,n1,n2)	求第一、二自由度分别为 n1、n2 的 F 分布在 x 处的概率密度值 y
y=ncx2pdf(x,n,delta)	求参数为 n,delta 的非中心卡方分布在 x 处的概率密度值 y
y=nctpdf(x,n,delta)	求参数为 n,delta 的非中心 t 分布在 x 处的概率密度值 y
y=ncfpdf(x,n1,n2,delta)	求参数为 n1,n2,delta 的非中心 F 分布在 x 处的概率密度值 y

【注】输入参数 x 可以是向量,此时输出 y 是同维数向量,下同.其他输入参数的意义和取值请查阅相关概率分布的数学定义.

(3) 累积概率分布函数(cdf)

函数及其调用格式	函数功能(分布参数意义同 pdf)
p=betacdf(x,a,b)	求 β 分布在 x 处的分布函数值 p
p=gamcdf(x,a,b)	求 γ 分布在 x 处的分布函数值 p
p=expcdf(x,lambda)	求指数分布在 x 处的分布函数值 p
p=normcdf(x,mu,sigma)	求正态分布在 x 处的分布函数值 p
p=logncdf(x,mu,sigma)	求对数正态分布在 x 处的分布函数值 p
p=unifcdf (x,a,b)	求均匀分布在 x 处的分布函数值 p
p=raylcdf(x,b)	求瑞利分布在 x 处的分布函数值 p
p=weibcdf(x,a,b)	求威布尔分布在 x 处的分布函数值 p
p=binocdf(x,n,P)	求二项分布在 x 处的分布函数值 p
p=poisscdf(x,lambda)	求泊松分布在 x 处的分布函数值 p
p=geocdf(x,P)	求几何分布在 x 处的分布函数值 p

p = hygecdf(x, m, k, n)	求超几何分布在 x 处的分布函数值 p
p = unicpdf(x, n)	求离散均匀分布在 x 处的分布函数值 p
p = nbincdf(x, R, P)	求负二项式分布在 x 处的分布函数值 p
p = chi2cdf(x, n)	求卡方分布在 x 处的分布函数值 p
p = tcdf(x, n)	求 t 分布在 x 处的分布函数值 p
p = fcdf(x, n1, n2)	求 F 分布在 x 处的分布函数值 p
p = ncx2cdf(x, n, delta)	求非中心卡方分布在 x 处的分布函数值 p
p = nctcdf(x, n, delta)	求非中心 t 分布在 x 处的分布函数值 p
p = ncfcdf(x, n1, n2, delta)	求非中心 F 分布在 x 处的分布函数值 p

【注】累积概率分布函数的数学定义为 $p = F(x) = P\{X \leqslant x\}$.

(4) 逆累积概率分布函数(inv)

函数及其调用格式	函数功能(分布参数意义同 pdf)
x = betainv(p, a, b)	求 β 分布的 p 分位点 x
x = gaminv(p, a, b)	求 γ 分布的 p 分位点 x
x = expinv(p, lambda)	求指数分布的 p 分位点 x
x = norminv(p, mu, sigma)	求正态分布的 p 分位点 x
x = logninv(p, mu, sigma)	求对数正态分布的 p 分位点 x
x = unifinv (p, a, b)	求均匀分布的 p 分位点 x
x = raylinv(p, b)	求瑞利分布的 p 分位点 x
x = weibinv(p, a, b)	求威布尔分布的 p 分位点 x
x = binoinv(p, n, P)	求二项分布的 p 分位点 x
x = poissinv(p, lambda)	求泊松分布的 p 分位点 x
x = geoinv(p, P)	求几何分布的 p 分位点 x
x = hygeinv(p, m, k, n)	求超几何分布的 p 分位点 x
x = unicinv(p, n)	求离散均匀分布的 p 分位点 x
x = nbininv(p, R, P)	求负二项式分布的 p 分位点 x
x = chi2inv(p, n)	求卡方分布的 p 分位点 x
p = tinv(p, n)	求 t 分布的 p 分位点 x
x = finv(p, n1, n2)	求 F 分布的 p 分位点 x
x = ncx2inv(p, n, delta)	求非中心卡方分布的 p 分位点 x
x = nctinv(p, n, delta)	求非中心 t 分布的 p 分位点 x
x = ncfinv(p, n1, n2, delta)	求非中心 F 分布的 p 分位点 x

【注】逆累积概率分布函数的数学定义为 $x = F^{-1}(p)$，即已知 $p = P\{X \leqslant x\}$，求 x.

(5) 均值和方差函数(stat)

函数及其调用格式	函数功能(分布参数意义同 pdf)
[M, V] = betastat (a, b)	求 β 分布的期望 M 和方差 V
[M, V] = gamstat (a, b)	求 γ 分布的期望 M 和方差 V
[M, V] = expstat (p, lambda)	求指数分布的期望 M 和方差 V
[M, V] = normstat(mu, sigma)	求正态分布的期望 M 和方差 V
[M, V] = lognstat (mu, sigma)	求对数正态分布的期望 M 和方差 V
[M, V] = unifstat (a, b)	求均匀分布的期望 M 和方差 V
[M, V] = raylstat (b)	求瑞利分布的期望 M 和方差 V
[M, V] = weibstat (a, b)	求威布尔分布的期望 M 和方差 V
[M, V] = binostat (n, P)	求二项分布的期望 M 和方差 V
[M, V] = poisstat (Lambda)	求泊松分布的期望 M 和方差 V
[M, V] = geostat (P)	求几何分布的期望 M 和方差 V
[M, V] = hygestat (m, k, n)	求超几何分布的期望 M 和方差 V
[M, V] = unidstat (n)	求离散均匀分布的期望 M 和方差 V
[M, V] = nbinstat (R, P)	求负二项式分布的期望 M 和方差 V
[M, V] = chi2stat (x, n)	求卡方分布的期望 M 和方差 V
[M, V] = tstat (n)	求 t 分布的期望 M 和方差 V
[M, V] = fstat (n1, n2)	求 F 分布的期望 M 和方差 V
[M, V] = ncx2stat (n, delta)	求非中心卡方分布的期望 M 和方差 V
[M, V] = nctstat (n, delta)	求非中心 t 分布的期望 M 和方差 V
[M, V] = ncfstat (n1, n2, delta)	求非中心 F 分布的期望 M 和方差 V

(6) 随机数产生函数(rnd)

函数及其调用格式	函数功能(分布参数意义同 pdf)
X = betarnd(a, b, r, c)	产生服从贝塔分布的 r 行 c 列随机数矩阵 X
X = gamrnd(a, b, r, c)	产生服从伽玛分布的 r 行 c 列随机数矩阵 X
X = exprnd(lambda, r, c)	产生服从指数分布的 r 行 c 列随机数矩阵 X
X = normrnd(mu, sigma, r, c)	产生服从正态分布的 r 行 c 列随机数矩阵 X
X = lognrnd(mu, sigma, r, c)	产生服从对数正态分布的 r 行 c 列随机数矩阵 X
X = unifrnd (a, b, r, c)	产生服从均匀分布的 r 行 c 列随机数矩阵 X
X = raylrnd(b, r, c)	产生服从瑞利分布的 r 行 c 列随机数矩阵 X
X = weibrnd(a, b, r, c)	产生服从威布尔分布的 r 行 c 列随机数矩阵 X
X = binornd(n, P, r, c)	产生服从二项分布的 r 行 c 列随机数矩阵 X

X=poissrnd(lambda,r,c)	产生服从泊松分布的 r 行 c 列随机数矩阵 X
X=geornd(P,r,c)	产生服从几何分布的 r 行 c 列随机数矩阵 X
X=hygernd(m,k,n,r,c)	产生服从超几何分布的 r 行 c 列随机数矩阵 X
X=unidrnd(n,r,c)	产生服从离散均匀分布的 r 行 c 列随机数矩阵 X
X=nbinrnd(R,P,r,c)	产生服从负二项式分布的 r 行 c 列随机数矩阵 X
X=chi2rnd(n,r,c)	产生服从卡方分布的 r 行 c 列随机数矩阵 X
X=trnd(n,r,c)	产生服从 t 分布的 r 行 c 列随机数矩阵 X
X=frnd(n1,n2,r,c)	产生服从 F 分布的 r 行 c 列随机数矩阵 X
X=ncx2rnd(n,delta,r,c)	产生服从非中心卡方分布的 r 行 c 列随机数矩阵 X
X=nctrnd(n,delta,r,c)	产生服从非中心 t 分布的 r 行 c 列随机数矩阵 X
X=ncfrnd(n1,n2,delta,r,c)	产生服从非中心 F 分布的 r 行 c 列随机数矩阵 X

二、关于常用统计量的 MATLAB 描述

统计量名称	函数及调用格式	函数说明
均值	M=mean(X)	计算向量 X 中各元素的算术平均值
	M=mean(X,dim)	计算向量 X 的指定维数 dim 内元素的算术平均值
	NM=nanmean(X)	计算向量 X 中除 NaN 外元素的算术平均值
	TM=trimmean(X,p)	计算向量 X 中元素的修正算术平均值,参数 p 表示所剔除的偏大和偏小数据的百分比
中值	ME=median(X)	返回向量 X 的中位数
	ME=nanmedian(X)	忽略 NaN 返回中位数
几何均值	GM=geomean(X)	计算向量 X 的几何平均值
调和均值	HM=harmmean(X)	计算向量 X 的调和平均值
p 分位数	xp=prctile(X,p)	计算 p 分位数. 其中输入参数 X 是数据向量,p 是取 0~100 的实数值;输出参数 xp 返回向量 X 的小于 p%(下侧)的分位点
最大值	MAX=max(X)	返回向量 X 的最大值
	MAX=nanmax(X)	返回忽略 NaN 的最大值
最小值	MIN=min(X)	返回向量 X 的最小值
	MIN=nanmin(X)	返回忽略 NaN 的最小值
排序	Y=sort(X)	返回向量 X 按由小到大排序后的向量

	[Y, I] = sort(X)	Y 为排序的结果, I 中元素表示 Y 中对应元素在 X 中位置
	Y = sort(X, dim)	在给定的 X 新的维数 dim 内排序. 若 X 元素为复数, 则按 |X| 排序
	Y = sortrows(X)	X 为矩阵, 返回矩阵 Y, Y 是按 X 的第 1 列由小到大、以行方式排序后生成的矩阵
	Y = sortrows(X, c)	按指定列 c 由小到大进行排序
	[Y, I] = sortrows(X, c)	Y 为排序的结果, I 表示 Y 中第 c 列元素在 X 中位置. 若 X 元素为复数, 则按 |X| 的大小排序
极差	R = range(X)	返回向量 X 中元素的最大值与最小值之差
方差	V = var(X)	计算向量 X 的方差
	V = var(X, 1)	计算向量 X 的二阶中心矩
	V = var(X, w)	计算向量 X 的以 w 为权重的方差
标准差	S = std(X, flag, dim)	计算向量 X 中维数为 dim 的元素的标准差值, 其中 flag = 0(默认)时, 置前因子为 1/(n − 1); 否则置前因子为 1/n. flag, dim 可缺省
	NS = nanstd(X)	若 X 为含有元素 NaN 的向量, 则返回除 NaN 外的元素的标准差
中心矩	B = moment(X, k)	计算向量 X 的 k 阶中心矩
峰度	KU = kurtosis(X)	计算向量 X 的峰度
偏度	SK = skewness(X)	计算向量 X 的偏斜度
协方差	C = cov(X)	计算向量 X 的协方差. 若 X 为矩阵, 返回 X 各列向量的协方差矩阵, 该协方差矩阵的对角线元素是 X 的各列的方差, 即 var(X) = diag(cov(X))
	C = cov(X, Y)	X, Y 为等长列向量, 等同于 cov([X Y])
相关系数	CR = corrcoef(X, Y)	计算列向量 X, Y 的相关系数
	CR = corrcoef (A)	返回矩阵 A 的列向量的相关系数矩阵

【注】上述函数中, 若 X 为矩阵, 则返回 X 中各列向量的函数值构成的行向量.

三、统计作图

(1) 正整数的频率表

【函数】tabulate

【格式】table = tabulate(X)

【说明】输入参数 X 为正整数构成的向量，返回矩阵 table 有 3 列：第 1 列为 X 的互异值，第 2 列为这些值的个数，第 3 列为这些值的频率.

(2) 经验累积分布函数图形

【函数】cdfplot

【格式】[h, stats] = cdfplot(X)

【说明】绘制样本 X(向量)的累积分布函数图形，并返回曲线的句柄 h 和若干样本特征值 stats.

(3) 最小二乘拟合直线

【函数】lsline

【格式】h = lsline

【说明】为数据散点图添加最小二乘拟合直线，h 为直线的句柄.

(4) 正态分布概率图形

【函数】normplot

【格式】h = normplot(X)

【说明】若 X 为向量，则显示正态分布概率图形；若 X 为矩阵，则显示每一列的正态分布概率图形. 返回绘图直线的句柄 h，样本数据在图中用"+"显示；如果数据来自正态分布，则图形显示为直线，而其他分布可能在图中产生弯曲.

(5) 威布尔(Weibull)概率图形

【函数】weibplot

【格式】h = weibplot(X)

【说明】若 X 为向量，则显示威布尔概率图形；若 X 为矩阵，则显示每一列的威布尔概率图形. 返回绘图直线的句柄 h. 绘制威布尔概率图形的目的是用图解法估计来自威布尔分布的数据 X，如果 X 是威布尔分布数据，其图形是直线的，否则图形中可能产生弯曲.

(6) 样本数据的盒图

【函数】boxplot

【格式】

① boxplot(X)

② boxplot(X, notch)

③ boxplot(X, notch, 'sym')

④ boxplot(X, notch, 'sym', vert)

⑤ boxplot(X, notch, 'sym', vert, whis)

【说明】

格式①产生矩阵 X 的每一列的盒图和"须"图，"须"是从盒的尾部延伸出来，并表示盒外数据长度的线，如果"须"的外面没有数据，则在"须"的底部有一个点.

格式②当 notch=1 时产生一凹盒图，notch=0 时产生一矩箱图.

格式③中 sym 表示图形符号，默认值为"+".

格式④当 vert=0 时生成水平盒图，vert=1 时生成竖直盒图(默认值 vert=1).

格式⑤中 whis 定义"须"图的长度，默认值为 1.5，若 whis=0，则 boxplot 函数通过绘制 sym 符号图来显示盒外的所有数据值.

(7) 给当前图形加一条参考线

【函数】refline

【格式】

① refline(slope, intercept)

② refline(slope)

【说明】

格式①中 slope 表示直线的斜率，intercept 表示截距.

格式②中 slope=[a b]，在图中加一条直线 y=b+ax.

(8) 在当前图形中加入一条多项式曲线

【函数】refcurve

【格式】h = refcurve(p)

【说明】在图中加入一条多项式曲线柄，p 为多项式系数向量，p=[p1, p2, p3, …, pn]，其中 p1 为最高幂项系数，h 为曲线的句柄.

(9) 样本的概率图形

【函数】capaplot

【格式】p = capaplot(data, specs)

【说明】data 为所给样本数据，specs 指定范围，p 表示在指定范围内的概率. 该函数返回来自估计分布的随机变量落在指定范围内的概率.

(10) 频数统计与频数直方图

【函数】hist

【格式】

① [N, A]=hist(data, nbins)

② hist(data, nbins)

【说明】data 是数据向量，nbins 指定数据分组数. 格式①可以完成各组数据频数的统计，N 返回各组的数据频数，A 返回各组数据的组中值. 格式②返回频数直方图.

(11) 附加有正态密度曲线的直方图

【函数】 histfit

【格式】histfit(data, nbins)

【说明】data为向量，返回直方图和正态曲线. nbins指定数据分组数，缺省时为data中数据个数的平方根.

(12) 在指定的界线之间画正态密度曲线

【函数】normspec

【格式】p = normspec(specs, mu, sigma)

【说明】specs指定界线，mu, sigma为正态分布的参数，p为样本落在上、下界之间的概率.

四、参数估计

(1) 几个特定分布的参数估计函数

函数及其调用格式	函数功能
[phat, pci] = binofit (x, n, alpha)	二项分布参数p的最大似然估计
[lambdahat, lambdaci] = poissfit (x, alpha)	泊松分布参数λ的最大似然估计
[muhat, sigmahat, muci, sigmaci] = normfit(x, alpha)	正态分布参数μ和σ的最大似然估计
[phat, pci] = betafit (x, alpha)	β分布参数a和b的最大似然估计
[ahat, bhat, aci, bci] = unifit(x, alpha)	均匀分布参数a和b的最大似然估计
[thatahat, thetaci] = expfit(x, alpha)	指数分布参数θ的最大似然估计
[phat, pci] = gamfit(x, alpha)	γ分布参数a和b的最大似然估计
[phat, pci] = weibfit(x, alpha)	威布尔分布参数a和b的最大似然估计

【注】上述各函数中输入参数x是样本数据向量，二项分布中输入参数n是试验次数；输出参数分两类。* hat是参数最大似然估计值，* ci是参数的显著性水平为alpha的置信区间，alpha的默认值为0.05.

(2) 通用极大似然估计函数

【函数】mle

【格式】[phat, pci] = mle('dist', data, alpha, n)

【说明】进行由dist的指定分布的分布参数的最大似然估计. data为样本数据向量，alpha为分布参数区间估计的显著性水平(缺省值为0.05)，n为试验总次数(仅用于二项分布)；返回参数phat和pci分别为分布参数的最大似然估计值和置信区间.

dist的取值包括：Beta, Bernoulli, Binomial, Discrete Uniform, Exponential, Extreme Value, Gamma, Geometric, Lognormal, Negative Binomial, Normal, Poisson, Rayleigh, Uniform, Weibull.

五、假设检验

(1) 正态变量均值的 U 检验法

【函数】ztest

【格式】[h, p, ci, zval] = ztest(x, m, sigma, alpha, tail)

【说明】检验的原假设为 $H_0: \mu = \mu_0 = m$. 输入参数 x 为样本向量；m 为待检验均值；sigma 为变量的标准差；alpha 为显著性水平(默认值 0.05)；tail 是备择假设说明：

若 tail=0, 表示备择假设为 $H_1: \mu \neq \mu_0 = m$(默认, 双边检验)；

若 tail=1, 表示备择假设为 $H_1: \mu > \mu_0 = m$(单边检验)；

若 tail=−1, 表示备择假设为 $H_1: \mu < \mu_0 = m$(单边检验).

输出参数 h 标示检验结论：

若 h=0, 表示在显著性水平 alpha 下, 不能拒绝原假设；

若 h=1, 表示在显著性水平 alpha 下, 可以拒绝原假设.

p 为检验的最小显著性概率；ci 为均值 μ 的 1−alpha 置信区间；zval 为统计量的值.

(2) 正态变量均值的 t 检验法

【函数】ttest

【格式】[h, p, ci, zval] = ztest(x, m, alpha, tail)

【说明】同 ztest.

(3) 两个正态变量均值差的 t 检验法

【函数】ttest2

【格式】[h, p, ci] = ttest2(x, y, alpha, tail)

【说明】检验的原假设为 $H_0: \mu_1 = \mu_2$. 输入参数 x, y 为两个变量的样本向量；其他参数同 ztest. 注意：tail = 0 表示备择假设为 $H_1: \mu_1 \neq \mu_2$; tail = 1 表示备择假设为 $H_1: \mu_1 > \mu_2$; tail = −1 表示备择假设为 $H_1: \mu_1 < \mu_2$.

(4) 连续变量分布形态的 Kolmogorov-Smirnov 检验法

【函数】kstest

【格式】[h, p, stat, cv] = kstest(x, cdf, alpha)

【说明】检验的原假设为变量 x 服从 cdf 指定的分布. 输入参数 x 为样本向量；cdf 为待检验的累积分布函数(cdf=[ ]时表示标准正态分布)；alpha 为显著性水平.

输出参数 h 标示检验结论；p 为检验的最小显著性概率；stat 为统计量的值；cv 为是否接受原假设的临界值.

(5) 两个连续变量分布一致性的 Kolmogorov-Smirnov 检验法

【函数】kstest2

【格式】[h, p, stat] = kstest2(x, y, alpha)

【说明】检验的原假设为两个变量具有相同的连续分布. 输入参数 x, y 为两个变量的样本向量；其他参数同 kstest.

(6) 两个变量分布一致性的秩和检验法

【函数】ranksum

【格式】[p, h, stat] = ranksum(x, y, alpha)

【说明】检验的原假设为两个变量具有相同的分布．参数同 kstest2．注意，stat 中包括：ranksum 为秩和统计量的值，zval 为计算 p 值使用的正态统计量的值．

(7) 两个变量中位数相等的符号秩检验法

【函数】signrank

【格式】[p, h, stat] = signrank(x, y, alpha)

【说明】检验的原假设为两个变量的样本中位数相等．参数同 kstest2．注意，stat 中包括：signrank 为符号秩统计量的值，zval 为计算 p 值使用的正态统计量的值．

(8) 两个变量中位数相等的符号检验法

【函数】signtest

【格式】[p, h, stat] = signtest(x, y, alpha)

【说明】同 signrank．

(9) 正态分布的拟合优度大样本检验法

【函数】jbtest

【格式】[h, p, stat, cv] = jbtest(x, alpha)

【说明】检验的原假设为变量 x 服从正态分布．参数同 kstest．

(10) 正态分布的拟合优度小样本检验法

【函数】lillietest

【格式】[h, p, stat, cv] = lillietest(x, alpha)

【说明】同 jbtest．

六、方差分析与回归分析初步

(1) 单因子方差分析

【函数】anova1

【格式】[p, anovatab, stats] = anova1(X, group, 'displayopt')

【说明】输入参数 X 是 r 个变量的 m 个样本观测值的 $m\times r$ 矩阵．group 是与 X 对应的表示 r 个变量的名字或意义的字符串数组，通常缺省使用．引用参数 displayopt 有两个状态 on 和 off，分别表示显示和隐藏方差分析表图形和 Box 图．

输出参数 p 为 X 的各列均值相等的最小显著性概率，p 的值越小，原假设越受置疑，表示这个因素对随机变量的影响是显著的．anovatab 和 stats 分别返回方差分析表和一个附加的统计数据结构，可以缺省．

(2) 双因子方差分析

【函数】anoval2

【格式】[p, table] = anova2(X, reps, 'displayopt')

【说明】输入矩阵 X 的行、列各表示一个因子，不同的行(列)表示该因子不同处理下的响应变量的观测值向量. 每一个“行与列的偶对”称为一个数据单元，如果各数据单元拥有多于一个的观测点，则参数 reps 声明每一个单元观测点的数目. 如在下面的矩阵中

$$\begin{matrix} A=1 & A=2 \end{matrix}$$

$$\begin{bmatrix} x_{111} & x_{112} \\ x_{121} & x_{122} \\ x_{211} & x_{212} \\ x_{221} & x_{222} \\ x_{311} & x_{312} \\ x_{321} & x_{322} \end{bmatrix} \begin{matrix} \left.\right\} B=1 \\ \left.\right\} B=2 \\ \left.\right\} B=3 \end{matrix}$$

行因子有三种不同处理，列因子有两种不同处理，每个数据单元不同数据标号(变动的下标)个数为 2，则 reps = 2(亦即每个数据单元行数与列数的较大者).

输出参数 p 是检验列、行及其交互作用均值相等的最小显著性概率(向量).

(3) 多元线性回归分析

【函数】regress

【格式】[b, bint, r, rint, stats] = regress(y, X, alpha)

【说明】用于 p 个自变量、一个因变量的线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim N(\boldsymbol{O},\ \sigma^2\boldsymbol{I})$ 的建模和模型评价. 其中，输入参数 X 表示 p 个自变量的 n 个观测值的 $n\times p$ 矩阵，y 表示因变量的 n 个观测值的 $n\times 1$ 向量，alpha 是显著性水平(可以缺省，此时默认为 0.05)；输出参数 b 返回的是模型系数(向量)β 的最小二乘估计值，bint 是 β 的 100(1 − alpha)% 置信区间，r 是模型拟合残差(向量)，rint 是模型拟合残差的 100(1 − alpha)% 置信区间，stats 包含可决系数 R^2 的值、方差分析的 F 统计量的值、方差分析的显著性概率 p 值和模型方差 σ^2 的估计值，bint、r、rint 和 stats 可以缺省.

(4) 逐步回归建模集成指令

【函数】stepwisefit

【格式】

[b, se, pval, inmodel, stats, nextstep, history] = stepwisefit(X, y, 'param1', value1, 'param2', value2, …)

【说明】用于 p 个自变量、一个因变量的线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim N(\boldsymbol{O},\ \sigma^2\boldsymbol{I})$ 的建模和模型评价. 其中，各参数的意义如下.

输入参数

X——p 个自变量的 n 个观测值的 $n \times p$ 矩阵.

y——因变量的 n 个观测值的 $n \times 1$ 向量.

'paramk'——第 k 个引用参数, valuek 是其取值, 通常可以缺省. 这里只介绍 3 个可能会用到的引用参数:

'penter'——设置回归方程显著性检验的显著性概率上限, 缺省设置为 0.05;

'premove'——设置回归方程显著性检验的显著性概率下限, 缺省设置为 0.10;

'display'——用来指明是否强制显示建模过程信息, 取值为'on'(显示, 缺省设置), 'off'(不显示).

输出参数

b——模型系数 .

se——模型系数的标准误差.

pval——显著性检验各个自变量的显著性概率.

inmodel——各个自变量在最终回归方程中地位的说明(1 表示在方程中, 0 表示不在方程中).

stats——一个构架数组, 包括:

source : 建模方法的说明, 'stepwisefit'表示逐步回归法;

dfe: 最优回归方程的剩余自由度;

df0: 最优回归方程的回归自由度;

SStotal: 最优回归方程的总偏差平方和;

SSresid: 最优回归方程的剩余平方和;

fstat: 最优回归方程的 F 统计量的值;

pval: 最优回归方程的显著性概率;

rmse: 最优回归方程的标准误差估计;

B: 模型系数;

SE: 模型系数的标准误差;

TSTAT: 每个自变量显著性检验的 T 统计量的值;

PVAL: 每个自变量显著性检验的显著性概率;

intercept: 常数项的点估计;

等等.

nextstep——对是否还有需要引入回归方程的自变量的说明(0 表示没有).

history——一个构架数组, 包括:

rmse: 每一步的模型标准误差估计;

df0: 每一步引入方程的变量个数;

in：记录了按相关系数绝对值大小逐步引入回归方程的变量的次序.

(5) 逐步回归建模交互式图形环境

【函数】stepwise

【格式】stepwise(X, y, inmodel, penter, premove)

【功能说明】创建多元线性回归分析的逐步回归法建模的交互式图形环境.

【参数说明】

X——p 元线性模型解释变量的 n 个观测值的 $n\times p$ 矩阵.

y——p 元线性模型因变量的 n 个观测值的 $n\times 1$ 向量.

inmodel——标量或向量(由 X 的列号构成)，用来指明最初引入回归方程的解释变量(缺省设置为空).

penter——模型检验的显著性水平上限值(缺省设置为 0.05).

premoveb——模型检验的显著性水平下限值(缺省设置为 0.10).

【交互式图形界面的说明】

窗口 Ⅰ Coefficients with Error Bars

绘出各个解释变量回归系数的估计，圆点表示点估计值，横线表示置信区间(有色线段表示 90%置信区间，黑色线段表示 95%置信区间). 窗口的右侧绘出回归系数的点估计值(Coeff)、显著性检验的 t 统计量的值(t-stet)和显著性概率 p 值(p-val).

窗口 Ⅱ Model History

该窗口绘出的圆点表示历次建模的模型标准差 σ 的估计.

两个窗口中间输出的是当前模型的有关信息，包括：

lntercept——模型截距(常数项)的估计；

RMSE——模型标准差 σ 的估计；

R-square——可决系数；

Adj-R-sq——校正的可决系数；

F——模型整体性检验的 F 统计量的值；

p——模型整体性检验的显著性概率.

窗口 Ⅰ 右侧的三个按钮：

Next Step ——在回归方程中按相关系数绝对值大小逐次引入解释变量，如无解释变量可引入时，按钮不可用.

All Steps ——直接给出“只进不出”方式建模的最终结果(注意，此时的回归方程未必是最优回归方程).

Export… ——选择向 Workspace 传输的计算结果(有关变量名可由用户自定义).

(6) 多元二次函数拟合建模

【函数】rstool

【格式】rstool(x, y, model, alpha)

【说明】多元二次函数为目标函数拟合建模. 输入参数 x 为 $n \times m$ 矩阵, y 为 n 维列向量, alpha 为显著性水平(缺省时设定为 0.05), model 由下列 4 个模型中选择 1 个(用字符串输入, 缺省时设定为线性模型):

linear(线性函数): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$;

purequadratic(纯二次函数): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{n} \beta_{ij} x_j^2$;

interaction(交叉二次函数): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leqslant j \neq k \leqslant m}^{n} \beta_{jk} x_j x_k$;

quadratic(完全二次函数): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leqslant j, k \leqslant m}^{n} \beta_{jk} x_j x_k$.

附录 C　正文中缺省的 M-文件

(1) normplot_1.m

```
%设定正态分布的分布参数μ和σ及绘图区域
clear
mu1=2.5;mu2=3;sigma1=0.5;sigma2=0.6;
x=(mu2-4*sigma2):0.01:(mu2+4*sigma2);
% 考察均值的影响
y1=normpdf(x,mu1,sigma1);
y2=normpdf(x,mu2,sigma1);
% 考察方差的影响
y3=normpdf(x,mu1,sigma1);
y4=normpdf(x,mu1,sigma2);
% 考察结果的可视化
subplot(1,2,1)
plot(x,y1,'-g',x,y2,'-b')
xlabel('\fontsize{12} μ1<μ2,σ1=σ2')
legend('μ1','μ2')
subplot(1,2,2)
plot(x,y3,'-g',x,y4,'-b')
xlabel('\fontsize{12} μ1=μ2,σ1<σ2')
legend('σ1','σ2')
```

(2) normplot_2.m

```
%(标准)正态分布密度曲线下的面积
clear,clf
X=linspace(-5,5,100);
Y=normpdf(X,0,1);
yy=normpdf([-3,-2,-1,0,1,2,3],0,1);
plot(X,Y,'k-',[0,0],[0,yy(4)],'c-.')
hold on
```

```
plot([-2,-2],[0,yy(2)],'m:',[2,2],[0,yy(6)],'m:',[-2,-0.5],[yy(6),
yy(6)],'m:',[0.5,2],[yy(6),yy(6)],'m:')
plot([-1,-1],[0,yy(3)],'g:',[1,1],[0,yy(5)],'g:',[-1,-0.5],[yy(5),yy
(5)],'g:',[0.5,1],[yy(5),yy(5)],'g:')
plot([-3,-3],[0,yy(1)],'b:',[3,3],[0,yy(7)],'b:',[-3,-0.5],[yy(7),yy
(7)],'b:',[0.5,3],[yy(7),yy(7)],'b:')
hold off
text(-0.5,yy(6)+0.005,'\fontsize{14}95.44%')
text(-0.5,yy(5)+0.005,'\fontsize{14}68.26%')
text(-0.5,yy(7)+0.005,'\fontsize{14}99.74%')
text(-3.2,-0.03,'\fontsize{10}μ-3σ')
text(-2.2,-0.03,'\fontsize{10}μ-2σ')
text(-1.2,-0.03,'\fontsize{10}μ-σ')
text(-0.05,-0.03,'\fontsize{10}μ')
text(0.8,-0.03,'\fontsize{10}μ+σ')
text(1.8,-0.03,'\fontsize{10}μ+2σ')
text(2.8,-0.03,'\fontsize{10}μ+3σ')
```

(3) chi2plot.m

```
%绘制不同自由度的卡方分布概率密度曲线
clear,clf
X=linspace(0,20,100);
Y1 = chi2pdf(X,1); %自由度等于1
Y2 = chi2pdf(X,3); %自由度等于3
Y3 = chi2pdf(X,6); %自由度等于6
plot(X,Y1,'-g',X,Y2,'-h',X,Y3,'-k')
%title('\fontsize{18}\fontname{华文新魏}不同自由度的{\chi}^2分布概率密度
曲线的比较')
text(0.6,0.6,'\fontsize{12}df:n=1')
text(2.6,0.2,'\fontsize{12}df:n=3')
text(8.6,0.09,'\fontsize{12}df:n=6')
%legend('df:n=1','df:n=3','df:n=6')
```

(4) tplot.m

```
%绘制 t 分布概率密度曲线
clear, clf
X = linspace( -4,4,100);
Y0 = normpdf(X,0,1); %标准正态分布
Y1 = tpdf(X,45); %自由度为 45
Y2 = tpdf(X,4); %自由度为 4
Y3 = tpdf(X,2); %自由度为 2
YY0 = normpdf(0,0,1);
plot(X,Y0,'.b',X,Y1,'-c',X,Y2,'-m',X,Y3,'-k',[0,0],[0,YY0],':r')
%title('\fontsize{18}\fontname{华文新魏}不同自由度的 t 分布概率密度曲线)
legend('N(0,1)','df:n=45','df:n=4','df:n=2')
```

(5) fplot_1.m

```
%绘制 F 分布概率密度曲线
clear, clf
X = linspace(0,6,100);
Y = fpdf(X,10,5); %自由度等于 10,5
plot(X,Y)
text(1.5,0.55,'\fontsize{14}df:n1=10,n2=5')
```

(6) fplot_2.m

```
%考察自由度对 F 密度曲线形态的影响
clear, clf
X = linspace(0,6,100);
Y11 = fpdf(X,100,10); %自由度等于 100,10
Y12 = fpdf(X,5,10); %自由度等于 5,10
Y21 = fpdf(X,10,100); %自由度等于 10,100
Y22 = fpdf(X,10,5); %自由度等于 10,5
subplot(2,1,1)
plot(X,Y11,X,Y12)
legend('df:100,10','df:5,10')
subplot(2,1,2)
plot(X,Y21,X,Y22)
legend('df:10,100','df:10,5')
```

(7) alphaplot.m

```
%α分位数示意图(标准正态分布,α=0.05)
clear
clf
data = normrnd (0,1,300,1);
xalpha1 = norminv (0.05,0,1);
xalpha2 = norminv (0.95,0,1);
xalpha3 = norminv (0.025,0,1);
xalpha4 = norminv (0.975,0,1);
subplot(3,1,1)
capaplot(data,[-inf,xalpha1]); axis([-3,3,0,0.45])
subplot(3,1,2)
capaplot(data,[xalpha2,inf]);axis([-3,3,0,0.45])
subplot(3,1,3)
capaplot(data,[-inf,xalpha3]);axis([-3,3,0,0.45])
hold on
capaplot(data,[xalpha4,inf]);axis([-3,3,0,0.45])
bold off
```

(8) stand.m

```
function X0 = stand(X)
% STAND 将数据矩阵 X 逐列进行标准化处理,输出标准化数据 X0
%语法
%X0 = stand(X)
%参数说明
% X—原始数据矩阵
% X0—标准化后的数据矩阵

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

zeros(size(X));
[nr,nx] = size(X);
for mk = 1:nr
```

```
X0(mk,:)=(X(mk,:)-mean(X))./std(X);
end
```

(9) bykpcr.m

```
function [W,C,T,U,P,R]=bykpcr(E0,F0)
%BYKPCR 提取 PLS 建模过程所有可能的主成分
%语法
% [W,C,T,U,P,R]=bykpcr(E0,F0)
%参数说明
% E0—自变量标准化的样本数据 n×p 矩阵
% F0—因变量标准化的样本数据 n×q 矩阵
% W—模型效应权重 p×rankE0 矩阵
% C—因变量权重 q×rankE0 矩阵
% T—自变量系统主成分得分 n×rankE0 矩阵
% U—因变量系统主成分得分 n×rankE0 矩阵
% P—模型效应载荷量 p×rankE0 矩阵
% R—因变量载荷量 q×rankE0 矩阵

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

A=rank(E0);
W=[];
C=[];
T=[];
U=[];
P=[];
R=[];
for byk=1:A
%提取主轴与主成分
EFFE=E0'*F0*F0'*E0;
FEEF=F0'*E0*E0'*F0;
options.tol=eps;
options.disp=0;
[w,LAMBDA]=eigs(EFFE,1,'lm',options);
```

```
[c,LAMBDA] = eigs(FEEF,1,'lm',options);
t1 = E0 * w;
u1 = F0 * c;
W = [W,w];
C = [C,c];
T = [T,t1];
U = [U,u1];
%计算残差
p1 = (E0' * t1)/norm(t1)^2;
E1 = E0 - t1 * p1'; E0 = E1;
r1 = (F0' * t1)/norm(t1)^2;
F1 = F0 - t1 * r1'; F0 = F1;
P = [P,p1];
R = [R,r1];
end
```

(10) plsra.m

```
function RA = plsra(T,R,F0,rankE0)
%PLSRA 求出主成分的累积复测定系数
%语法
% RA = plsra(T,R,F0,rankE0)
%参数说明
% T —自变量系统主成分得分 n×rankE0 矩阵
% R —因变量载荷量 q×rankE0 矩阵
% F0 —因变量标准化的样本数据 n×q 矩阵
% rankE0 —plspcr 提取的主成分个数

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

RAAA = [];
for byk = 1:rankE0
RAbyk = sum(norm(T(:,byk)).^2 * norm(R(:,byk)).^2)./(norm(F0))^2;
RAAA = [RAAA,RAbyk];
end
```

```
RA = cumsum(RAAA) ;
```

(11) plsrd.m

```
function [Rdx,RdX,RdXt,Rdy,RdY,RdYt] = plsrd(E0,F0,T,h)
%函数功能
% PLSRD 分析主成分解释能力
%语法
% [Rdx,RdX,RdXt,Rdy,RdY,RdYt] = plsrd(E0,F0,T,h)
%参数说明
% E0 —标准化后的自变量数据
% F0 —标准化后的因变量数据
% T —自变量系统主成分得分 n×rankE0 矩阵
% h —用于建模或希望进行解释能力分析的主成分个数
% Rdx —各主成分对于某自变量的解释能力
% RdX —各主成分对自变量组的解释能力
% RdXt —全部主成分对自变量组的累计解释能力
% Rdy —各主成分对于某因变量的解释能力
% RdY —各主成分对因变量组的解释能力
% RdYt —全部主成分对因变量组的累计解释能力

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

% - - 成分对自变量解释能力分析 - -
[nr,nx] = size(E0);
[nr,ny] = size(F0);
Rdx = zeros(nx,h);
t1 = zeros(nr,1);
x1 = zeros(nr,1);
for xj = 1:nx
for ti = 1:h
t1 = T(:,ti);
x1 = E0(:,xj);
cc = (corrcoef(t1,x1)).^2;
Rdx(xj,ti) = cc(1,2);
```

```
end
end
Rdx;
RdX = sum(Rdx)./nx ;
RdXt = sum(RdX);
% --成分对因变量解释能力分析--
Rdy = zeros(ny,h);
y1 = zeros(nr,1);
for yj = 1:ny
for ti = 1:h
t1 = T(:,ti);
y1 = F0(:,yj);
rr = (corrcoef(t1,y1)).^2;
Rdy(yj,ti) = rr(1,2);
end
end
Rdy;
RdY = sum(Rdy)./ny;
RdYt = sum(RdY);
```

(12) plsutcor.m

```
function cr = plsutcor(U,T)
% PLSUTCOR 绘制 t1/u1 图,并给出二者的相关系数
%语法
% cr = plsutcor(U,T)
%参数说明
% U —因变量提取的成分
% T —自变量提取的成分
% cr —自变量与因变量的相关系数

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

u1 = U(:,1);
t1 = T(:,1);
```

```
ut=[u1,t1];
cr=corrcoef(ut)
plot(t1,u1,'o')
lsline
title('t1/u1 图')
xlabel('t1')
ylabel('u1')
```

(13) pls.m

```
function SCOEFF=pls(h,p,W,P,R)
% PLS 求偏最小二乘回归方程的系数
%语法
% SCOEFF=pls(h,p,W,P,R)
%参数说明
% h 一用于建模的主成分个数
% p 一自变量个数
% W 一模型效应权重 p×rankE0 矩阵
% P 一模型效应载荷量 p×rankE0 矩阵
% R 一因变量载荷量 q×rankE0 矩阵
% SCOEFF 一偏最小二乘回归方程的系数 p×q 矩阵

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

for byk=1:h
if byk==1
WX(:,byk)=W(:,byk);
SCOEFF=WX(:,byk)*R(:,byk)';
else
I=eye(p);
ww=eye(p);
for bykbyk=1:byk-1
ww=ww*(I-W(:,bykbyk)*P(:,bykbyk)');
end
WX(:,hyk)=ww*W(:,byk);
```

```
end
SCOEFF=WX(:,byk)*R(:,byk)';
end
```

(14) plsiscoeff.m

```
function [COEFF,INTERCEP]=plsiscoeff(X,Y,B)
% PLSISCOEFF 标准化回归系数逆标准化处理,输出原始自变量对因变量的回归系数及常数项
%语法
% [COEFF,INTERCEP]=plsiscoeff(X,Y,B)
%参数说明
% X —原始自变量数据
% Y —原始因变量数据
% B —标准化变量回归方程的系数
% COEFF —原始变量回归方程的系数
% INTERCEP —原始变量回归方程的常数项

% 编写于 2007 年 5 月 18 日,修改于 2007 年 11 月 12 日

[xrow,xcol]=size(X);
[yrow,ycol]=size(Y);
for i=1:ycol
bykCOEFF(:,i)=B(:,i).*std(Y(:,i));
end
for j=1:ycol
for i=1:xcol
COEFF(i,j)=bykCOEFF(i,j)/std(X(:,i));
end
end
INTERCEP=mean(Y)-(mean(X)*COEFF);
```

(15) plsvip.m

```
function VIP=plsvip(W,RdY,RdYt,h)
% PLSVIP 进行变量投影重要性分析
```

```
%语法
% VIP=plsvip(W,RdY,RdYt,h)
%参数说明
% W 一自变量提取的主轴值
% RdY 一各成分对因变量组的解释精度
% RdYt 一全部成分对因变量组的解释精度
% h 一用于建模的主成分个数
% VIP 一变量投影重要性指标

% 编写于2007年5月18日,修改于2007年11月12日

[nx,wk]=size(W);
VIP=zeros(1,nx);
for j=1:nx
for hh=1:h
Whj=W(j,hh);
tvip(hh)=RdY(hh)*Whj.^2;
end
S_tvip=sum(tvip);
VIP(j)=sqrt((nx/RdYt)*S-tvip);
end
bar(YIP,'c')
title('变量投影重要性 VIP 图')
```

参考文献

[1] 茆诗松,程依明,濮晓龙. 概率论与数理统计教程[M]. 北京:高等教育出版社,2004.

[2] 庄楚强,何春雄. 应用数理统计基础[M]. 广州:华南理工大学出版社,2006.

[3] 王梓坤. 概率论基础及其应用[M]. 北京:北京师范大学出版社,1996.

[4] 李涛,贺勇军,刘志俭. MATLAB 工具箱应用指南:应用数学篇[M]. 北京:电子工业出版社,2000.

[5] 周明,李长虹,雷虎民. MATLAB 图形技术:绘图及图形用户接口[M]. 西安:西北工业大学出版社,1999.

[6] 陆璇. 数理统计基础[M]. 北京:清华大学出版社, 1998.

[7] 复旦大学. 概率论:第 2 册　数理统计[M]. 北京:人民教育出版社, 1979.

[8] 刘金兰. 管理统计学[M]. 天津:天津大学出版社,2007.

[9] 杨虎,刘琼荪,钟波. 数理统计[M]. 北京:高等教育出版社,2004.

[10] 杨振海,张忠占. 应用数理统计[M]. 北京:北京工业大学出版社,2005.

[11] 王岩,隋思莲,王爱青. 数理统计与 MATLAB 工程数据分析[M]. 北京:清华大学出版社,2006.

[12] 高惠璇. 应用多元统计分析[M]. 北京:北京大学出版社,2005.

[13] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京:国防工业出版社,1999.

[14] R L 奥特,M 朗格内克. 统计学方法与数据分析引论[M]. 张忠占,等译. 北京:科学出版社,2003.

[15] 赵选民. 实验设计方法[M]. 北京:科学出版社,2006.